Master Thesis

Efficient modeling of head movements and dynamic scenes in virtual acoustics

Author
Tobias Grämer

Advisors
Martin Müller
Steven Schimmel

USZ, *Laboratory of Experimental Audiology*
ETHZ, *Institute for Biomedical Engineering*
June 2, 2010

Abstract

Algorithms for the improvement of speech intelligibility in hearing prostheses can degrade the spatial quality of the sound signal. To investigate the influence on distance perception and localization of such algorithms, a system to virtually render arbitrary static acoustical scenes has been developed. In this master thesis, the existing virtual acoustics system has been extended to present more realistic dynamic scenes. The system is able as well to compensate for the head movements of the test subject.

Subjective listening tests were conducted to evaluate the extended system. Static sources remain stable even in the case of fast head movements, the externalization of sound sources is improved compared to the existing system and simulated sound sources are nearly indistinguishable from real sound sources.

Acknowledgments

I would like to thank Martin Müller and Steven Schimmel for their inputs, support and their great help. I am also grateful to Prof. Norbert Dillier for having given me the opportunity of doing my thesis work at the University Hospital Zurich in the Laboratory for Experimental Audiology. I would like further to thank the whole LEA research group for the pleasant atmosphere during the many lunches and coffee breaks. A special thank goes to Andrea Kegel who always had time to listen to my problems and who thought me the basics of statistics. Last but not least, I would like to thank to all my patient test subjects, Michael Büchler, Peter Derleth, Markus Hofbauer, Andrea Kegel, WaiKong Lai, Martin Müller, Juliane Raether, and Steven Schimmel. Without them, this work would not have been possible.

Preface

This Master Thesis is part of my graduate study at the Department of Information Technology and Electrical Engineering (D-ITET) at the Swiss Federal Institute of Technology (ETH).

I certify that this Master Thesis, and the research to which it refers, is the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referring practices of the discipline.

Tobias Grämer

Author:	Tobias Grämer	graemeto@ee.ethz.ch
Advisors:	Martin Müller	martin.mueller@orl.usz.ch
	Steven Schimmel	steven.schimmel@usz.ch
Professors:	Norbert Dillier	norbert.dillier@orl.usz.ch
	Peter Bösiger	boesiger@biomed.ee.ethz.ch

Contents

List of Figures

List of Tables

Abbreviations

ANOVA	:	Analysis of Variance
BRIR	:	Binaural Room Impulse Response
CIC	:	Completely in the Canal
CPU	:	Central Processing Unit
DSP	:	Digital Signal Processor
FFT	:	Fast Fourier Transform
HRIR	:	Head-Related Impulse Response
HRTF	:	Head-Related Transfer Function
ITD	:	Interaural Time Difference
ITE	:	In The Ear
ILD	:	Interaural Level Difference
MAA	:	Minimum Audible Angle
MAMA	:	Minimum Audible Movement Angle
MLS	:	Maximum Length Sequences
MSE	:	Mean Squared Error
RMS	:	Root Mean Square
VBAP	:	Vector Base Amplitude Panning

Chapter 1

Introduction

1.1 Motivation

To investigate the spatial quality of hearing prostheses, a system to virtually render arbitrary static acoustical scenes has been developed. A virtual acoustics system provides an easy, flexible and controllable environment to test several hearing aid algorithms (or parts of it) and to quantify their influence on localization and distance perception. This implies that the virtual acoustics system is working perfect, i.e. that it is able to generate sound scenes which are indistinguishable from reality. The existing system [1] almost fulfils this requirement, but only for static scenes. It is unable to render moving sound sources or to handle head movements of test subjects. If a subject moves his head, the virtual acoustic scene that is presented to the subjects moves together with the subject, rather than staying fixed in world coordinates.

Previous research [2] has shown that such head and source movements facilitate correct and accurate sound localization judgments, reduce front-back confusions and allow better source elevation recognition. Similarly, it was found that the rate of front-back confusions is somewhat higher when virtual acoustics are involved compared to normal listening conditions. It is therefore desirable to extend the virtual acoustic system to include subject head movement and dynamic scenes.

1.2 Expected Problems and Possible Solutions

To simulate a static acoustic scene, the virtual acoustics system generates a binaural room impulse response (BRIR) for a given source and receiver position and orientation. To simulate head movements and dynamic scenes, this BRIR must be constantly updated to reflect the new position of the sound source and orientation of the receiver. For the intended purpose, the sound source updates would come form a pre-described sound source trajectory, and receiver orientation updates would come in real-time from a motion tracker device. The main limitations which need to be considered:

1. It is difficult if not impossible to generate the full BRIR in real-time with low latency.

This problem is usually addressed by separating the BRIR into several parts, where each part is updated at a rate high enough so that the overall BRIR remains perceptually convincing, but that is low enough so that the entire BRIR is available in real-time and with low latency.

2. Head-related transfer functions (HRTFs) are typically measured for a limited number of sound source positions on a sphere around the subject.

   The spatial sampling of HRTFs is usually dense enough so that convolving direct sound and surface reflections from arbitrary directions with their nearest-neighbour HRTF measurements will maintain perceptual accuracy. However, the spatial sampling may be too coarse to render small head movements and smooth sound source movements in a perceptually convincing way. Typical solutions are to describe the measured HRTFs in a model, such that HRTFs for intermediate directions can be synthesized from the model, or to interpolate HRTFs for intermediate directions from adjacent measured HRTFs.

1.3 Similar Systems

There are several types of virtual acoustics systems today, but they are either not freely available or do not satisfy our requirements. A broad class of systems are virtual reality systems which simulate not only the acoustical world but also the visual world. An early and prominent system of this type was "CAVE – Audio Visual Experience Automatic Virtual Environment" [3]. A similar system was "DIVA – An Integrated System for Virtual Audio Reality" [4]. A recent project with an emphasis on sound field reproduction is a system of the RWTH Aachen University, developed by Tobias Lentz et al. [3]. The system uses four loudspeakers instead of a headphone to generate a binaural acoustical scene. The audio system realizes the computation of the several tasks on dedicated machines that are interconnected by a network. The purpose of that type of systems is to compute the impulse response of a given room setup in real time, which is then used to augment the images of a virtual reality system with a plausible, but not necessarily physically correct acoustic room impression.

Another class of simulators are room acoustics software programs such as Odeon [5]. These programs are designed to simulate the acoustics of geometrically complex rooms such as churches, theaters and concert halls for accurate prediction and diagnosis of room acoustic properties. They require precise specification of the room geometry, their algorithms are optimized for the computation of room acoustics properties rather than for a room impulse response, and they are not necessarily tuned for speed.

1.4 Outline

This report is divided into the following Chapters:

Chapter 2 gives a short introduction into the theory of room acoustics, room simulation, sound localization and the measurement and interpolation of head-related transfer functions.

Chapter 3 explains the virtual acoustics system in more detail and describes how it was extended to account for head movements and dynamic scenes.

Chapter 4 describes how the modified virtual acoustics system was evaluated and summarizes the results of the listening tests

Chapter 5 subsumes the main results of this project and provides an outlook for further research in this area.

Chapter 2

Theory

This chapter gives an overview of some theoretical aspects and concepts which are important for this work. This includes some basic theory of human auditory localization, room acoustics and HRTF interpolation.

2.1 Coordinate Systems

In addition to the standard 3-dimensional Cartesian coordinate system, there are two other coordinate systems used in this work; they facilitate the handling of HRTFs or head movements.

2.1.1 Euler Angles

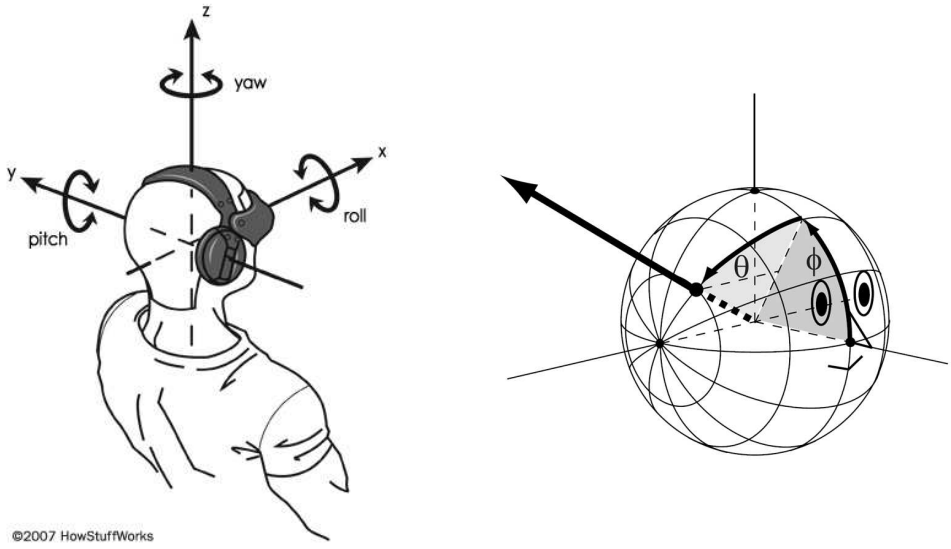
The Euler Angles are not a coordinate system in the strict sense but rely on the standard Cartesian coordinate system. The are used to describe the *orientation* of a rigid body in 3-dimensional Euclidean space.

In this work, a variant of proper Euler Angles is used, namely the yaw, pitch and roll notation. Yaw, pitch and roll each describe a rotation around one axis of the Cartesian coordinate system as illustrated in Figure 2.1 a).

2.1.2 Interaural-Polar Coordinate System

This spherical coordinate system is shown in Figure 2.1 b). The polar axis is the line behind the ears. The vertical plane that bisects the head left and right is called the median plane. The direction of a ray from the origin to a sound source (or vice versa) is specified by two angles, the azimuth angle θ and the elevation angle ϕ. In interaural-polar coordinates, the azimuth θ is the angle between the ray and the median plane. The elevation ϕ is the polar angle, specifying the rotation around the interaural axis.

A number of things follows from this definition. First, the median plane is defined by $\theta = 0°$. The azimuth is always between 0° and 360°, with increasing values for a clockwise rotation (90° is on the right, 270° on the left). In general, a surface of constant azimuth is a cone. A surface of constant elevation is a plane. In particular, the surface where $\phi = 0°$ is the horizontal plane. While



(a) Yaw Pitch Roll angles [39] (b) The interaural-polar coordinate system [8]

Figure 2.1: Coordinate systems

the range of azimuth is full 360°, the range of elevation is only 180°. -90° is the position below the head and 90° the overhead position [8].

2.2 Room Acoustics

This section deals with room acoustics and provides a short summary about the various influences a sound wave experiences when traveling from a sound source to a receiver in an enclosed space, based on a paper from Schimmel [7].

First, the sound source has a specific radiation characteristic, that is, the intensity of emitted sound varies in dependency of direction and frequency. Having left the source, the intensity of the sound decreases with the square of the distance it has traveled. The air absorption will also lower the intensity of the sound. If a sound wave hits a boundary of a room, a part of it is absorbed, a part of it is reflected specularly and a part is reflected diffusely. The amount of absorption, specular reflections, and diffuse reflections are frequency dependent properties of the room surface. Sound is also absorbed and reflected by objects in the room, and diffracted around edges. When a sound wave reaches a receiver (a microphone or an eardrum), the finally recorded or perceived sound depends on the receiver's directional sensitivity and frequency characteristics.

The so-called *room impulse response* is a "fingerprint" of the acoustical properties of a room. It represents all the sound propagation paths from a source to a receiver. It consists of several perceptually relevant components, as shown in Figure 2.2: direct sound, early specular reflections and the reverberant tail. The direct sound is the part of the sound which reaches the receiver without hitting a surface. Perceptually, it is the most important part since it reaches the receiver first and - in most cases - with the highest intensity. It is primarily

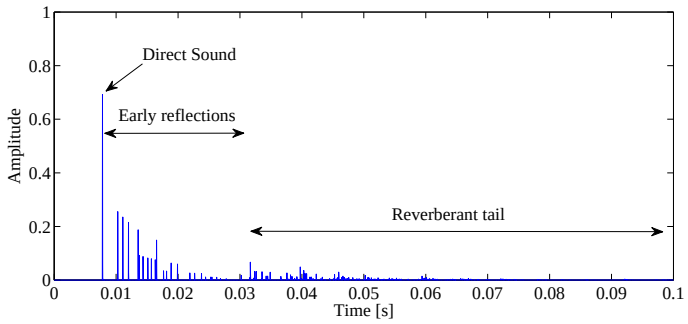


Figure 2.2: Components of a room impulse response

responsible for the localization of a sound source. The sound waves which are reflected a few times from surfaces are denoted as early reflections. They arrive at the receiver roughly within 50 ms after the direct sound. Usually, their arrival rate is low enough such that the human auditory system can relate them to a specific sound source. The last part of the impulse response, the reverberant tail, includes all the propagation paths that reflect off many surfaces before reaching the receiver. Their arrival rate is so high that the auditory system can not resolve them individually, but rather integrates them temporally and spatially into a combined percept. The reverberant tail characterizes the surface absorption and size of the room, and the ratio of direct sound to reverberation energy translates into a sense of distance of the sound source. A *dry* or *anechoic* room refers to a highly absorbent room with a fast decaying reverberant tail. A *live* or *reverberant* room describes a highly reverberant room with a long impulse response.

2.3 Room Modelisation and Simulation

The purpose of a room simulation software is to generate a room impulse response of a given room which represents the acoustical properties of this specific room. Compared to the measurement of an impulse response in a real room, the software based solution is usually much faster, cheaper and also more flexible, because it allows to generate impulse responses for a bunch of source and receiver positions.

To generate the impulse response of a certain room, this room has to be specified in terms of geometry and surface properties, that is, frequency dependent absorption and diffusity coefficients of the walls. Furthermore, a source and a receiver has to be defined, that is, the position, orientation and also the acoustical properties (directional, frequency-dependent radiation characteristic and sensitivity, respectively). The software then computes the impulse response from the source to the receiver.

In this project, the software ROOMSIM is used [7]. It is a portable, fast and flexible to use room simulator for “shoebox” rooms. The simulation is perceptually accurate because it models both specular and diffuse surface reflections. The first are simulated with the virtual image source method, the

latter with the diffuse rain algorithms. For details, the reader is referred to Schimmel [7].

2.4 Perception and Localization

Hearing is not a purely mechanical phenomenon of wave propagation, but is also a sensory and perceptual event. When a person hears something, that something arrives at the ear as a mechanical sound wave traveling through the air, but within the ear it is transformed into neural action potentials. These nerve pulses then travel to the brain where they are perceived. The study of subjective human perception of sounds is called psychoacoustics. Some topics of this broad field which are important for this work are shortly described in this section, for a more detailed description, we refer to the literature, e.g. [27].

2.4.1 Sound Localization

Sound localization refers to a listener's ability to identify the location or origin of a sound in direction and distance. A first theory on sound localization based on binaural cues was proposed by Lord Rayleigh in 1907 already. In his duplex theory, he supposed that interaural time differences (ITDs) and interaural level differences (ILDs) allow the localization of sound sources in the free field. For low frequencies, the ITDs provide the dominant cue whereas for high frequencies the ILDs are more important for the localizationThe transition from ITDs to ILDs occurs gradually at $f_c \approx 1.5$ kHz.

The duplex theory is based on the observation that, for frequencies below f_c, the dimensions of the head are smaller than the half wavelength of the sound waves. Therefore, the auditory system can determine phase delays between both ears unambiguously. In contrast, ILDs are very low in this frequency range, so that a precise evaluation of the input direction is nearly impossible on the basis of level differences only. For frequencies above f_c, the dimensions of the head are greater than the length of the sound waves. An unambiguous determination of the input direction based on interaural phases is not possible at these frequencies. However, the interaural level differences become bigger, and these level differences are evaluated by the auditory system. Despite its simplicity, the duplex theory has been verified in a broad range of discrimination experiments for a wide variety of stimuli [37].

The mechanisms described above explain how the azimutal position of a sound source can be identified, but they cannot be used to determine its elevation angle. Due to the symmetry of the head and the ears, there are directions with equal ITDs and ILDs which form a so-called *cone of confusion*. In particular, the ITDs and ILDs provide no information if a source is located ahead of a listener or behind him, the result is an uncertainty about the direction of the sound, a so-called *front-back confusion*. To reduce the amount of front-back confusions, additional cues are evaluated by the auditory system. The human outer ear, i.e. the structures of the pinna and the external ear canal, form direction-selective filters which provide additional information about a source position. However, these cues are weaker than the ITDs and ILDs and therefore,

not all front-back confusions can be resolved, especially in difficult situations with a low signal to noise ratio (SNR). Wallach suggested in 1940 [13], that head or source movements can resolve front-back confusions, a hypothesis which is supported by more recent studies [2], [14].

Studies about the ability to localize objects in motion revealed some effects not present in static scenes. In an initial approach, Perrott and Musicant [15] presented a sound stimulus that rotated around the subject's head. Subjects estimated the horizontal position of the sound source at the moment when the sound started (i.e., its onset position) and the moment when it ceased (i.e., its offset position). In the results, both the apparent onset and offset positions were mainly displaced in the direction of motion. Later studies were conducted with varying stimuli, source trajectories, velocities and with various test equipment and answer devices, respectively.

2.4.2 Precision of Localization

The spatial resolution of the human auditory system was investigated under a wide range of conditions with various stimuli. For static scenes, minimum audible angle (MAA) thresholds are determined and minimum audible movement angles (MAMA) for moving sources, respectively [31]. In the static case, listeners usually had to localize a sound source relative to another source. Sound was first played from one source and, after a short pause, played from the other source. In MAMA experiments, listeners have typically been asked to discriminate between directions of motion or to discriminate between a stationary and a moving sound source. For both MAAs as well as MAMAs, localization is most precise around 0° azimuth and becomes worse with increasing azimuth [28]. The MAA is $\sim 1°$ around 0° azimuth and increases up to $\sim 15°$ for 90° azimuth. The increase is, however, not linear, with a MAA below $\sim 5°$ for 60° azimuth. MAMAs are as much as several times larger than static minimum audible angles measured under comparable conditions. Near 0° azimuth, the MAMA is $\sim 8°$ [28], increasing to $\sim$ 10-15° around 60° azimuth and with a maximum of more than 30° at 90° azimuth, depending on the source speed. It was also shown that over a broad range of source velocities, the MAMA increases linearly with increasing source velocity [29]. Smallest MAMAs were measured at a source speed of 1.8 °/sec, for source speeds below that threshold, the MAMAs increase again. It can be concluded, that "thresholds associated with the detection of motion (MAMA) and with binaural spatial resolution (MAA) are probably independent" [30].

2.4.3 Head-Related Transfer Functions

Head-related transfer functions, sometimes also denoted as head-related impulse responses (HRIRs) model the acoustic path from a source located in the free field to the eardrum, in an anechoic environment. HRTFs describe how a sound is filtered by the diffraction and reflection properties of the head, pinna, shoulders and torso of a subject, before the sound reaches the eardrum. This implies that HRTFs differ for every subject. When measured precisely, HRTFs

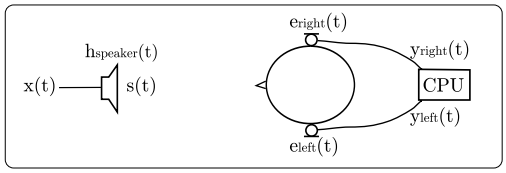


Figure 2.3: HRTF measurement setup [1]

allow to reproduce the same sound waves at the eardrum as if they would come from the corresponding position in space [1]. If sound is presented over headphones, it is often perceived in the head. When the same sound is filtered with the individual HRTF of that subject, it is possible to create the impression of an external sound source. A good externalization is only possible with individual, precisely measured HRTFs, an improved perception is also possible with generic HRTFs, but there are large inter-individual differences. In summary, HRTFs are crucial for a virtual acoustics system to simulate perceptually convincing scenes.

2.5 HRTF Measurement

A common method to measure impulse responses in the time domain is to use a maximum-length sequence (MLS). MLS are pseudo random binary sequences which are spectrally flat, that means their energy is equally distributed along the frequency spectrum. MLS additionally have the property that their autocorrelation function yields an impulse signal and the cross-correlation function of a system's response to an MLS with the MLS itself is the system's impulse response. For a detailed description of impulse response measurement using MLS, please refer to Heutschi [6].

Figure 2.3 shows the setup of the system used to measure HRIRs as employed by Müller [1]. A MLS $x(t)$ is generated, played trough a loudspeaker and two signals, $y_l(t)$ and $y_r(t)$ are recorded by the left and right microphones, located inside the ear canal. Before arriving at the microphone, the signals are changed by the measurement hardware and the room. These undesired effects have to be compensated for. First, the loudspeaker characteristics modify $x(t)$. The signal $s(t)$ then travels trough the room before arriving at the microphone position in the left and right ears. Room reflections, traveling delay, air attenuation, the individual characteristics of the outer ear and the resonances in the ear canal modify $s(t)$. Eventually, the recorded signals $y_l(t)$ and $y_r(t)$ are affected by the individual characteristics of the microphones. All those influences except the loudspeaker characteristics can be compensated for. For a detailed description, please refer to Müller [1]. The final HRTF can be derived from the HRIR by simply appliying the Fourier Transform to the measured and compensated HRIR.

2.6 HRTF Interpolation

2.6.1 Overview

HRTFs are typically measured for a limited number of sound source positions on a sphere around the subject. The resulting spatial sampling is usually dense enough such that convolving direct sound and surface reflections from arbitrary directions with their nearest-neighbour HRTF measurements will maintain perceptual accuracy. However, it may be too coarse to render small head movements and smooth sound source movements perceptually convincing. Typical solutions are to describe the measured HRTFs in a model, such that HRTFs for intermediate directions can be synthesized from the model, or to interpolate HRTFs for intermediate directions from adjacent measured HRTFs.

In this work, HRTFs are measured for 12 equal spaced positions in the horizontal plane. We decided to use interpolation to generate HRTFs for all the other positions. In general, interpolation could be done in the time or in the frequency domain and one could use one of several standard interpolation methods like linear interpolation, sinc or spline interpolation. However, it is not sufficient in the case of HRTFs to assess a certain interpolation scheme only from a technical point of view (i.e. using an interpolation method which leads to a low mean square error (MSE)), an authentic subjective impression is also important.

2.6.2 Literature Review

Several interpolation schemes are discussed in literature, a good summary of them can be found in a paper of Ajdler [11], the most important ones are also listed in this section.

The most simple and straightforward method is linear interpolation in the time domain. Nearest neighbour HRTFs are used to obtain HRTFs at any position in between. More elaborated interpolation techniques like polynomial or spline interpolation are of course also possible, but all those methods show a poor performance in a MSE sense as well as in subjective listening tests.

A decomposition of HRTFs in a minimum phase and an all-pass function was first proposed by Kistler [17] and picked up again by Kulkarni [16]. It was an important step towards the interpolation of time-aligned HRTFs. Since the minimum phase components have a minimum phase lag, phase delay, and group delay for a given magnitude, they are optimally aligned in time. This idea of alignment in time has been further refined in other papers. Indeed, it has been shown that the performance of interpolation in the time or frequency domain can be improved by compensating HRTFs prior to interpolation according to the time of arrival of sound [18], [19]. That is, the HRTFs are time aligned and interpolation is carried out on the time-aligned HRTFs. In order to achieve sub-sample precision in the time alignment, the time of arrival itself is also interpolated. For the interpolation of the time-aligned HRTFs, standard interpolation techniques like linear, spline and sinc interpolation were compared and the best results are obtained using linear interpolation [19].

Other methods that were considered are the decomposition in a limited number of basis functions with corresponding weighting factors depending on azimuthal and elevation angles. The basis functions can be determined by principal components analysis [17], independent component analysis [20] or spatial feature extraction and regularization [21]. With the measured HRTFs at known positions, the weighting factors can be derived. Finally, these weighting factors are interpolated to synthesize a HRTF at any position.

Ajdler et al. [11] recently proposed two new techniques and compared them with simple sinc interpolation in the time domain on the one hand and with the interpolation of time-aligned HRTFs on the other hand. For both methods, HRTFs are divided in a low-frequency and in a high-frequency part. The cut-off frequency is derived from the spatial Nyquist theorem which indicates that below a certain frequency (depending on the spacing between consecutive samples), there is enough information available for precise interpolation. However, it is important to consider that the spatial Nyquist theorem is derived under the assumption of free field conditions and does not take into account head shadowing nor diffraction. In both proposed methods, they apply spatial sinc interpolation for the low-frequency parts of the HRTFs. For the high frequency parts of the HRTF, the proposed methods differ: Their first approach is to interpolate time aligned HRTFs as described in [19]. The second approach is new. In the high frequency parts, the interpolation is carried in the complex temporal envelope domain in subbands. The interpolated subbands are obtained by restoring the carrier after interpolating the complex envelopes. They compare the different approaches with numerical simulations and assess the performance. Their proposed methods perform better than pure interpolation of time aligned HRTFs. However, the performance assessment is based on the interpolation of models and MIT KEMAR data [22], but not on real HRTFs. Furthermore, their only performance measure is the mean square error, they don't provide any subjective listening tests.

2.6.3 Time Aligned HRTF Interpolation

According to the spatial sampling theorem [11], the angular sampling frequency needs to satisfy

$$l_{\theta_S} > 2|\omega_{max}|\frac{d}{2c} \approx 2|2\pi f_{max}|\frac{0.09}{c},$$

where l_{θ_S} denotes the angular sampling frequency (i.e. the number of measurement points in the horizontal plane), d the distance between the two microphones and c the speed of sound. For an average adult human, $d \approx 0.18m$. Without any of the described methods, we would need a spacing of at least 13.6° for the precise interpolation of HRTFs sampled at 16 kHz ($f_{max} = 8000$ Hz). For a spacing of 30°, precise interpolation is only possible up to $f_{max} = 3640$ Hz, higher frequencies suffer from aliasing. To ease this problem, we decided to use the well-studied method of interpolating time aligned HRTFs.

Time Alignment

In a first step, the measured HRTFs are upsampled by a factor of 10 in order to achieve sub-sample precision in the estimate of the time delay. Then, the HRTFs are time aligned with respect to the direct sound of the first HRTF (0° azimuth). The left and the right HRTFs are aligned separately. The delay of the HRTFs is calculated with the cross correlation. Cross correlation provides excellent time delay estimation for broadband signals and for narrow band signals in the low-frequency range. However, for high frequency narrow band signals, it produces multiple ambiguous peaks [12]. Since HRTFs have a large bandwidth, we can use the method anyway. We calculated the maximum of the normalized cross correlation in the time domain. Let us denote $h_1[k]$ as the first HRTF (0° azimuth) and $h_2[k]$ as the second HRTF (any azimuth). Then, the normalized cross correlation is defined as

$$x[k] = \frac{1}{n-1} \sum_{m=-n+1}^{n-1} \frac{(h_1[m] - \overline{h_1})(h_2[k+m] - \overline{h_2})}{\sigma_{h_1}\sigma_{h_2}}$$

where n is the length, $\overline{h}$ the mean and σ_h the standard deviation of an HRTF. Then, we are looking for the maximum of the cross correlation,

$$\max_k(x[k]).$$

The position k of the maximum corresponds to the delay t relative to the first HRTF. To align the HRTFs, we are using standard sinc interpolation:

$$h_{aligned}[k] = h[k] * \mathrm{sinc}[k-t]$$

where sinc[.] is the normalized sinc function defined as

$$\mathrm{sinc}(z) = \begin{cases} \frac{\sin(\pi z)}{\pi z} & z \neq 0 \\ 1 & z = 0. \end{cases}$$

The time aligned HRTFs are interpolated in the time domain by means of monotone piecewise cubic interpolation [9], [10].

Monotone Piecewise Cubic Interpolation

Let az_i, $i = 1, 2, \ldots, n_{az}$ be the azimuth of the ith measured HRTF. Let Δaz_i be the ith subinterval (the angular distance between two measured HRTFs):

$$\Delta az_i = az_{i+1} - az_i.$$

Then, the first *divided difference*, $\Delta h_i[k]$, is given by

$$\Delta h_i[k] = \frac{h_{i+1}[k] - h_i[k]}{\Delta az_i}.$$

Let d_i denote the slope of the interpolant at az_i. If $sgn(\Delta h_i[k]) \neq sgn(\Delta h_{i-1}[k])$ or if $\Delta h_i[k] = 0$ or if $\Delta h_{i-1}[k] = 0$, then $d_i[k]$ is given by:

$$d_i[k] = 0.$$

This is the case if $h_i[k]$ is a local minimum or maximum. Otherwise, $d_i[k]$ is given by:

$$d_i[k] = \frac{1}{\frac{1}{2}\left(\frac{1}{\Delta h_{i-1}[k]} + \frac{1}{\Delta h_i[k]}\right)}$$

The HRTF at azimuth az between az_i and az_{i+1} is then given by:

$$h(az)[k] = h_i[k]H_1(az) + h_{i+1}[k]H_2(az) + d_i[k]H_3(az) + d_{i+1}[k]H_4(az),$$

where $H_1(az) = \phi((az_{i+1} - az)/\Delta az_i)$, $H_2(az) = \phi((az - az_i)/\Delta az_i)$, $H_3(az) = -\Delta az_i \psi((az_{i+1} - az)/\Delta az_i)$, $H_4(az) = \Delta az_i \psi((az - az_i)/\Delta az_i)$, where $\phi(z) = 3z^2 - 2z^3$, and $\psi(z) = z^3 - z^2$.

After the interpolation, the original time delay is restored and the HRTFs are downsampled to the original sampling frequency.

Evaluation of Interpolation

Figure 2.4 and 2.5, respectively shows an example of an interpolated HRTF. With simple linear interpolation in the time domain, there would be destructive interference due to the different times of arrival.

The quality of the interpolation was also assessed in a MSE sense. The MSE at an angular position θ_0 is defined as

$$MSE(\theta_0) = 10 \log_{10} \frac{\sum_{n=0}^{T} (h(\theta_0, n) - h_e(\theta_0, n))^2}{\sum_{n=0}^{T} h^2(\theta_0, n)},$$

where T stands for the number of time samples of the HRTF [11]. Figure 2.6 shows the average MSE as a function of the angular position. The HRTFs were taken from the MIT KEMAR database [22], which provides a spatial sampling frequency of 72 (5° spacing) in the horizontal plane. We have considered the case of interpolation of HRTFs every 20° to obtain HRTFs every 10°. The MSE averaged over all interpolated positions is -16.2 dB, which coincides with the results of Ajdler.

Informal subjective listening tests (cf. section 3.2.3) indicate that there were no audible artefacts. Therefore, we kept this method although the technique proposed by Ajdler performs slightly better in a MSE sense.

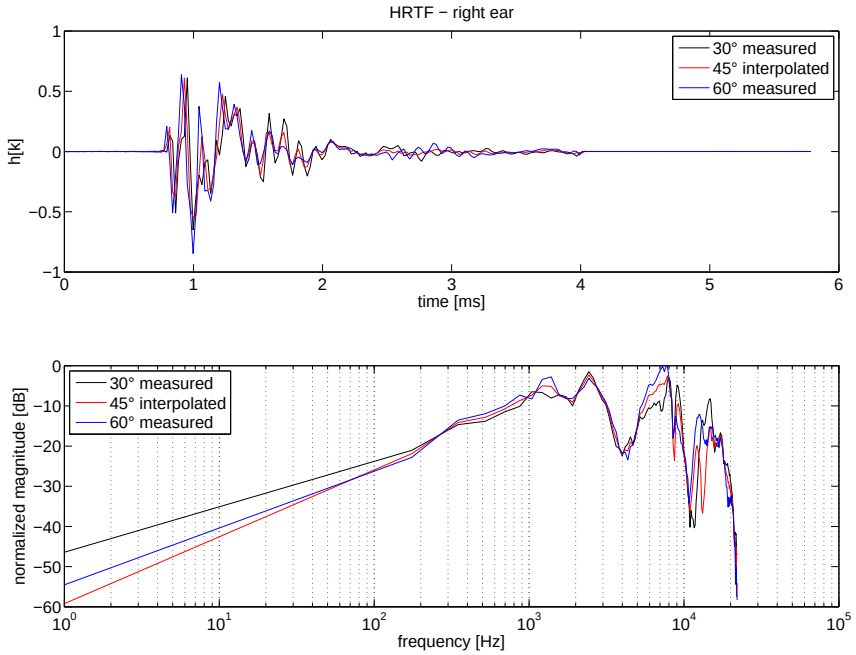


Figure 2.4: Example of a real, interpolated HRTF at azimuths of 30° (measured), 45° (interpolated) and 60° (measured)

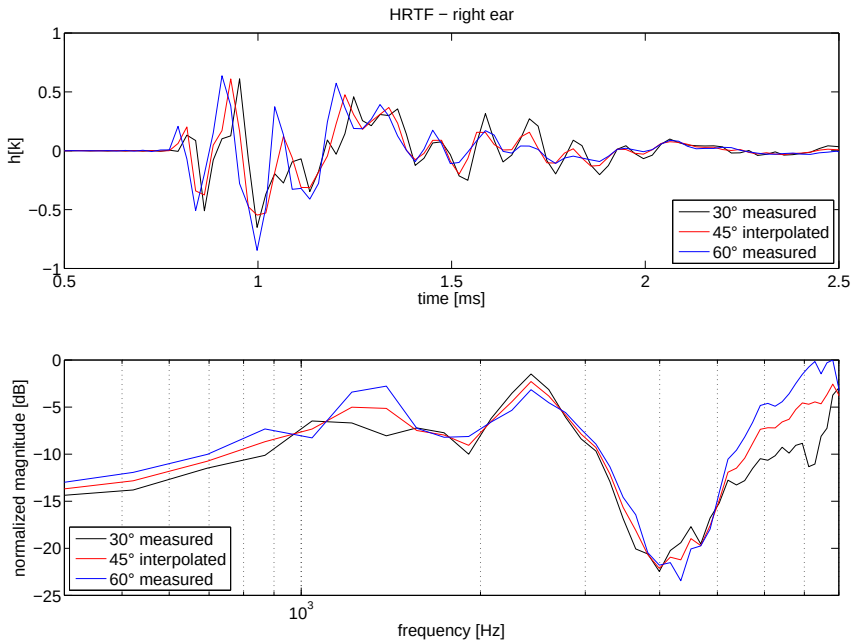


Figure 2.5: Detailed view of the same interpolated HRTF. The time domain plot is truncated to the range from 0.5 ms to 2.5 ms; the frequency domain plot shows the frequencies from 400 Hz up to 8 kHz.

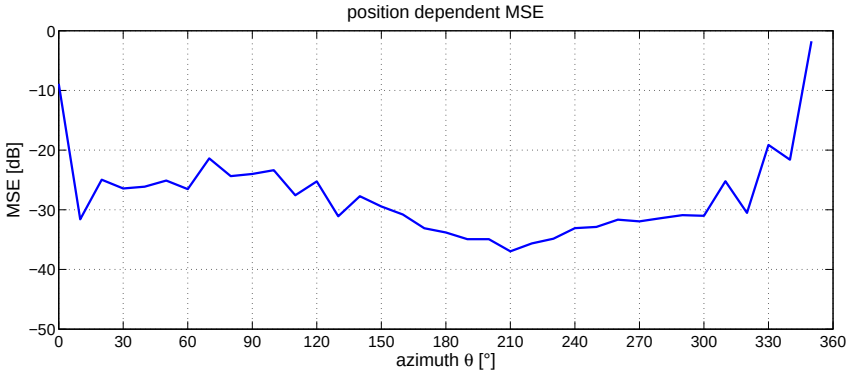


Figure 2.6: Average MSE error in the case of a spacing of 20° in the database.

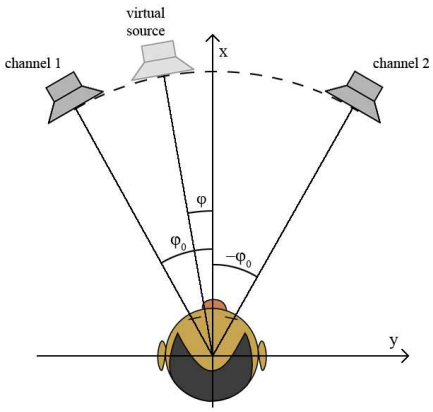


Figure 2.7: Amplitude panning: two-channel stereophonic configuration [26]

2.7 Amplitude Panning

Amplitude panning is used to generate virtual sound sources with two or more loudspeakers. In the simple amplitude panning method, two loudspeakers radiate coherent signals which may have different amplitudes. The listener perceives an illusion of a single auditory event (virtual sound source, phantom sound source), which can be placed on a two-dimensional sector defined by locations of the loudspeakers and the listener by controlling the signal amplitudes of the loudspeakers. This is also known as stereophonic playback. In this work, we use vector base amplitude panning (VBAP) to generate virtual, moving sound sources on a circle around a listener in the horizontal plane. This section provides a short summary of some important aspects of VBAP based on a paper from Pulkki, for a detailed description we refer to the original paper [26].

The situation is illustrated in Figure 2.7. The amplitudes of the two loudspeaker signals are controlled with gain factors g_1 and g_2, respectively. In our setup, the angle $\varphi_0 = 15°$. The direction of the virtual source is dependent on

the relation of the amplitudes of the emanating signals. If the virtual source is moving and its loudness should be constant, the gain factors that control the channel levels have to be normalized. The sound power can be set to a constant value C, whereby the following approximation can be stated:

$$g_1^2 + g_2^2 = C. \tag{2.1}$$

The directional perception of a virtual sound source produced by amplitude panning follows approximately the stereophonic law of sines:

$$\frac{\sin\varphi}{\sin\varphi_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{2.2}$$

where $0° < \varphi_0 < 90°$, $-\varphi_0 \leq \varphi \leq \varphi_0$, and $g_1, g_2 \in [0, 1]$. In Eq. 2.2, φ represents the angle between the axis and the direction of the virtual source; $\pm\varphi_0$ is the angle between the x axis and the loudspeakers.

This two-channel stereophonic loudspeaker configuration can be reformulated as a two-dimensional vector base. The base is defined by unit-length vectors $\mathbf{l}_1 = [\ l_{11} \quad l_{12}\ ]^T$ and $\mathbf{l}_2 = [\ l_{21} \quad l_{22}\ ]^T$, which are pointing toward loudspeakers 1 and 2, respectively. The unit-length vector $\mathbf{p} = [\ p_1 \quad p_2\ ]^T$, which points toward the virtual source, can be treated as a linear combination of loudspeaker vectors,

$$\mathbf{p} = g_1\mathbf{l}_1 + g_2\mathbf{l}_2. \tag{2.3}$$

We may write the equation in matrix form,

$$\mathbf{p}^T = \mathbf{g}\mathbf{L}_{12}$$

where $\mathbf{g} = [\ g_1 \quad g_2\ ]$ and $\mathbf{L}_{12} = [\ \mathbf{l}_1 \quad \mathbf{l}_2\ ]^T$. This equation can be solved if $\mathbf{L}_{12}^{-1}$ exists,

$$\mathbf{g} = \mathbf{p}^T\mathbf{L}_{12}^{-1} = [\ p_1 \quad p_2\ ] \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1}. \tag{2.4}$$

When $\varphi_0 \neq 45°$, the gain factors have to be normalized using the equation

$$\mathbf{g}^{scaled} = \frac{\sqrt{C}\mathbf{g}}{\sqrt{g_1^2 + g_2^2}}.$$

Now gain factors $\mathbf{g}^{scaled}$ satisfy Eq. 2.1. The extension for a system with more than two loudspeakers is straightforward, since this setup can be decomposed in a set of pair-wise loudspeakers where the discussed VBAP can be applied.

Chapter 3

Implementation

This chapter describes the existing virtual acoustics system and explains in detail how it was extended to account for head movements and dynamic scenes. The hardware and the test environment are introduced and the limitations of the system are discussed.

3.1 Hardware

3.1.1 Head-Tracker

For the compensation of head movements, these movements have to be measured by a sensor which will also be denoted as the "head-tracker". We used a commercial sensor, the Xsens MTx 3DOF Orientation Tracker. The MTx consists of rate of turn sensors, accelerometers and magnetometers. Table 3.1 lists the most important specifications and Figure 3.1 shows a picture of the sensor.

The actual performance of the sensor does not comply with these specifications, in particular the static sensor accuracy is much worse than 1°, which limits the performance of the whole system. Chapter 4.1 deals with the achieved performance and the resulting limitations of the head-tracker in our setup.

The motion tracking sensor is mounted on the subject's head with a cap, Figure 3.2 shows the very beautiful, fashionable, custom-tailored solution. It is

Static accuracy (roll/pitch)	$< 0.5°$
Static accuracy (heading)[a]	$< 1°$
Dynamic accuracy[b]	2° RMS
Angular resolution[c]	0.05°
Maximum update rate, onboard processing	120 Hz
Maximum update rate, external processing	512 Hz

[a] under condition of a stabilized Xsens sensor fusion algorithm
[b] 1 σ standard deviation of zero-mean angular random walk, may depend on type of motion
[c] in homogeneous magnetic environment

Table 3.1: Technical Specification of the head-tracker

Figure 3.1: Xsens MTx motion tracking sensor

Figure 3.2: Head-tracker mounted on a dummy head

important that the cable is also attached to the subject's head such that the sensor is not moved relative to the head for arbitrary head movements. The chosen solution fulfils this condition.

The position and orientation of the sensor on the head may vary, this is compensated by the Xsens sensor fusion algorithm that processes the raw sensor data. Although it is possible to access to the raw sensor data, we always used the output from the sensor fusion algorithm in the "Euler" format (cf. section 2.1). The algorithm has a few parameters which can be fine-tuned:

- **Weighting Factor**
  Indicates how much the sensor data from the magnetometer should be weighted relative to the accelerometer data. A number of 1 indicates the magnetometer data is considered equal to the accelerometer data and this should be the default value. A number of 0.0 will completely disregard any data from the magnetometers, otherwise valid range is <0.1 ; 10].

- **Filter Gain**
  The gain is the most important tweaking option. Very roughly the gain equals the "cross-over" frequency of the sensor fusion algorithm in Hertz. For example, a value of 1 for the gain means, more or less, that frequency components of the calculated orientation vector exceeding 1 Hz will be determined by the rate of turn sensors and components below 1 Hz will be determined by the accelerometers and magnetometers. Valid values are larger than 0.01 and lower than 50, i.e. <0.01 .. 50], some values may lead to unstable operation of the algorithm under certain conditions. The recommended default value of the gain is 1.

- **Adapt to Magnetic Disturbances** Large amounts of ferrous material (iron, nickel and cobalt but not e.g. aluminum and most stainless steels)

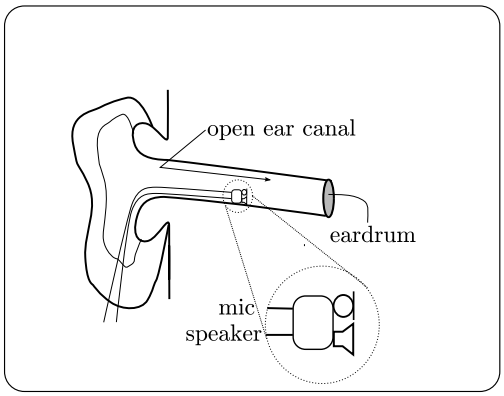


Figure 3.3: Open ITE micro speaker prototype [1]

will disturb the homogeneous earth magnetic field used as a reference by the sensor. The sensitivity of the system to such disturbance can be significantly reduced by an advanced sensor fusion algorithm setting called AMD (Adapt to Magnetic Disturbances). The default or "normal" operating mode should however be with this option turned OFF as drift around the vertical (yaw/heading) will occur over time.

3.1.2 Playback Devices

Open ITE Micro Speaker

For the best possible sound reproduction, we developed a special playback and recording device, in the following denoted as open ITE speaker or simply speaker [1]. The term ITE means "in the ear". It is schematically depicted in Figure 3.3. It consists of an individually designed pair of miniature microphone and speaker mounted on an open shell of completely in the canal (CIC) hearing aids, such that the sound from outside can pass trough the device. Compared to headphones, this micro speaker system has the following advantages: 1) the close location of the microphone to the speaker ($\approx 2mm$) ensures that the sound is played at the same location where it is measured. 2) The system sits always at the same location in the ear canal. Repeated measurements and playback show minimal differences. 3) The ear canal is open during playback. This allows a more natural and comfortable sound reproduction.

Loudspeakers

For the assessment of the virtual acoustics system, we compared simulated scenes with the corresponding real scenes where sound is played over loudspeakers. We used twelve "Genelec Active Monitor Model 1029A" loudspeakers. In the following, they are simply referred as loudspeaker.

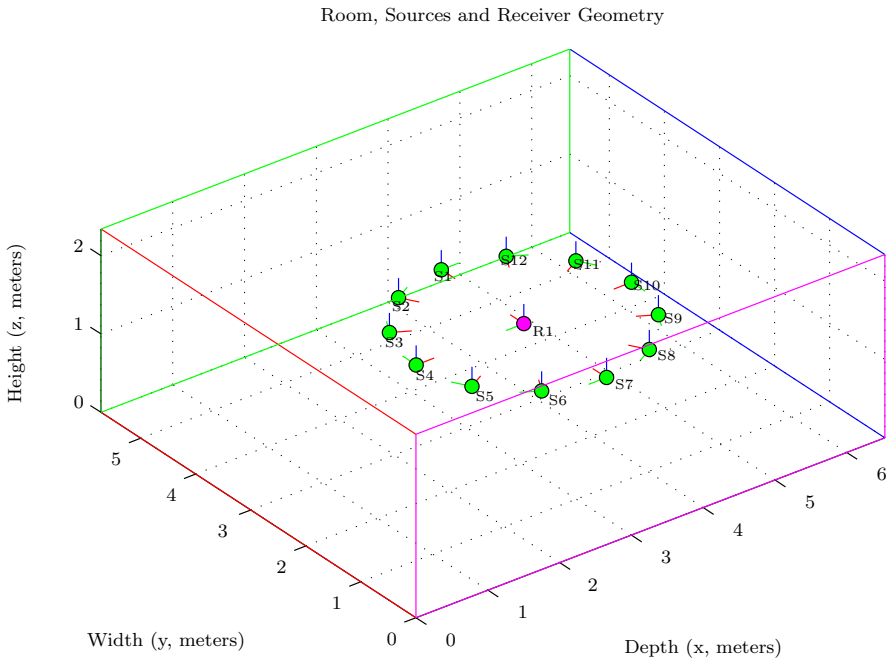


Figure 3.4: Room simulation setup [1]

3.1.3 Other Hardware

All audio signals and the head-tracker data were processed by a standard PC equipped with an Intel Core2Duo processor running at 2.66 GHz, 4 GB of RAM and custom software running with Matlab version 7.5.0.342 (R2007b) under Windows XP SP3. Two soundcards RME Hammerfall DSP Multiface II were used for D/A conversion.

A Norsonic 118 sound level meter was used for the measurement of sound pressure levels. They were all measured with an A-weighting [dB(A)] and with the time constant "slow" (1 sec).

3.1.4 Room Setup and Test Environment

The room where all listening tests took place was an acoustically treated shoebox-type room with octave-band reverberation times (T_{60}) shown in Table 3.2. The room was 6.53 meters large, 5.72 wide and 2.34 high. The receiver, i.e. the head of a listener was at position (3.69, 2.85, 1.15) in the centre of a circle of 12 loudspeakers with an angular spacing of 30° and a radius of 1.5 meters as shown in Figure 3.4. HRTFs were measured for those 12 positions and interpolated with a resolution of 1° for all positions between the loudspeakers. For the filtering of reflections outside the horizontal plane, we used anechoic KEMAR HRTFs [22], which were also interpolated to obtain a 1° spatial resolution.

frequency [Hz]	125	250	500	1000	2000	4000	8000
$T_{60}[ms]$	230	270	270	210	230	300	300

Table 3.2: Measured octave band reverberation times [1]

3.2 Virtual Representation

3.2.1 Existing System

To simulate an acoustical scene, we use the software ROOMSIM (cf. section 2.3)in combination with a custom-made Matlab-based program for all the filtering and the sound output. The loudspeaker is modeled as a source with frequency independent radiation characteristics of a subcardioid, given by:

$$I(\theta, \phi) = 0.7 + \cos(\theta)\cos(\phi)$$

where θ denotes the azimuth, ϕ the elevation and I the intensity in [dB] radiated from the loudspeaker in that direction. Figure 3.5 shows a two-dimensional subcardioid. The frequency dependent absorption and diffusity coefficients of the room were determined empirically to match the measured binaural room impulse response. The receiver characteristics are defined by the measured and interpolated HRTFs. With all those parameters, ROOMSIM can generate a BRIR for specified source and receiver positions. The resulting impulse responses were eventually calibrated to compensate the ear canal resonance and microphone, which is done by inverse filtering with the respective impulse response. In a frequency domain notation, the final impulse response from the source to the receiver for one ear, $H_{s,r}$, is given by

$$H_{s,r}(f) = \frac{H_s(f)R_{s,r}(f)H_r(f)}{H_{calib}(f)}$$

where $R_{s,r}$ is the room impulse response from the source to the receiver, representing room reflections, traveling delay and air attenuation. H_s denotes the source characteristics and H_r the subject's individual HRTF which models the effects of the torso, the shoulders, the head and the pinnae. H_{calib} is the transfer function from the open ITE speaker to the microphone right next to it, it describes the strong resonance that occur in the ear canal when playing the sound with the open ITE speakers. H_{calib} is measured by means of the same MLS correlation procedure that is used to measure the HRTFs (cf. section 2.5).

When we have calculated the full impulse response, we can simply convolve an audio signal with $h(t)$ and play the resulting signal with the open ITE speakers.

3.2.2 From Static to Dynamic Scenes

With a room simulation software which is able to generate BRIRs for arbitrary source and receiver positions and with individual (interpolated) HRTFs in the whole horizontal plane, we have a good basis to generate also dynamic scenes.

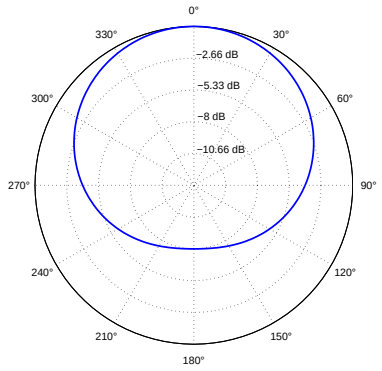


Figure 3.5: Loudspeaker radiation characteristics: Subcardioid

The difference between static and dynamic scenes is that in a dynamic scene, the BRIR is changing continuously in dependence of the source and the receiver position and orientation. In a digital world, continuous processes are usually approximated by discretization, that is, updating the process in very short time intervals. In our case, this means that we need to be able to update our BRIR at a high rate, either fixed or depending on movements of the source and/or the receiver.

The generation of a full BRIR is a computationally intensive task which is very hard to implement in real-time with low latency. To overcome this problem, one could use dedicated, powerful hardware and optimize the software for this specific platform. A real-time DSP based system would probably be the most promising approach, since DSPs are special processors designed for signal processing and to meet real-time requirements. An optimization and adaption of the roomsim software to benefit from the possibilities of todays multi-core processors in a standard PC could also be an option. This would require to parallelize the calculations and to make use of advanced instruction sets, that is, to use assembly language to fully exploit the power of a specific processor. A similar approach is the use of a graphics card as a processor, also known as General Purpose Computation on Graphics Processing Unit (GPGPU), a fast growing technology. However, GPGPU is more suitable for tasks which can be calculated offline, because the latency of data transfers from and to the GPU is relative high. Common to all those approaches is the loss of flexibility and portability which motivated the development of the ROOMSIM software used in this project. Moving to another platform or system would contradict this philosophy of portability.

Another way to meet the real-time requirements with the existing system is to reduce the computational complexity. This is usually done by separating the BRIR into several parts, where each part is updated at a rate that is high enough so that the overall BRIR remains perceptually convincing, but that is low enough so that the entire BRIR is available in real-time and with low latency. This strategy requires a lot of subjective listening tests to assess the

effects of a somewhat pruned BRIR. The two ideas, optimizing of the software and the reduction of computational complexity, respectively can of course also be combined.

A third possibility is to calculate the BRIRs offline for all source and receiver positions and orientations, which requires a spatial discretization in addition to the temporal discretization. Since we have to consider all source positions, source orientations, receiver positions and receiver orientations, each of them with 3 degrees of freedom, this results in a total of 12 degrees of freedom. Although memory is cheap today, it is clear that this approach is a dead end if we want to render arbitrary movements.

Since the first approach, using dedicated hardware and optimizing the calculations, is hard to integrate into the existing system and also contradicts the portability paradigm, we decided to build a system without any additional hardware (except the head-tracker), that is compatible to the existing Matlab-based implementation. The real-time filtering of the audio signal (cf. section 3.2.4) is already using most of the resources of our computer, therefore we did not try to generate the BRIR in real-time. For the intended purpose of the system, it is sufficient to restrict source movements to a circle around the receiver and receiver movements to the yaw axis. That way, from the 12 degrees of freedom, 10 are eliminated. With a spatial sampling of 1°, the remaining two degrees of freedom require to calculate $360^2 = 129600$ BRIRs which is still too much. Our measurement room is almost symmetric, which allowed us to further reduce the number of pre-computed impulse responses. The details are described in section 3.2.4.

Other virtual acoustics systems described in the literature (cf. section 1.3) provide some evidence for the required spatial resolution and the maximum overall processing delay which is acceptable. Lentz et al. [3] suggest that "[...] a filter change every 1-2 degrees is necessary. In order to be precise for almost all possible rotational velocities, we consider a timing interval for a recalculation every 10-20 milliseconds as mandatory. As a consequence, the block size should not be bigger than 512 samples as this limits the minimal possible update time to 11.6 milliseconds at a 44.1 kHz sampling rate."

3.2.3 Dynamic Scenes without Head Movements

This section describes the offline rendering of dynamic scenes without support for head movements. Without head movements, there are no real-time requirements so that the processing is straightforward. Nevertheless, we could gain some experiences which allowed us to estimate what temporal and spatial resolution is required for a smooth impression.

For a sound source which is moving at a constant speed on a circle around the listener, one could either define a spatial resolution or a temporal resolution, the other property is then determined by the speed of the source. We set a fixed temporal resolution and restricted also the spatial resolution, that is, we used fixed nearest-neighbour positions of the actual sound source to generate the impulse response. Additionally, we tried linear interpolation of the two nearest neighbour positions. The processing of the audio signal was done in a block-wise

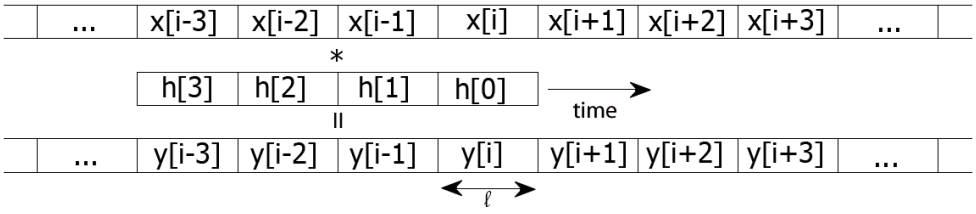


Figure 3.6: Block-wise signal processing. $x[i]$ denotes an input signal block i, $y[i]$ an output signal block i and h is the impulse response divided into N blocks. In this example, $N = 4$. All blocks are of the same length l.

manner as illustrated in Figure 3.6.

The processing of moving sources implies a source position dependent impulse response h_θ. Since the trajectory of the moving source is known in advance, the position dependency can also be formulated as a time dependency. The output signal in the case of a moving source is obtained by:

$$y[k] = \sum_{n=0}^{N-1} \sum_{m=nl}^{(n+1)(l-1)} x[k-m]h_\theta[m] \tag{3.1}$$

where the indices k and m address single samples of the signal and the impulse response, respectively, x denotes the input signal, N the number of blocks the impulse response is divided into and l is the block length. h_θ is the position dependent impulse response, it is calculated the same way as in the existing system for static scenes (cf. 3.2.1). Every input block is filtered with the impulse response corresponding to the source position at the time when this very block is played. The inner sum is nothing else than the convolution of an input signal block with a block from the impulse response. Reformulating equation 3.1 in terms of block processing yields:

$$y[i] = \sum_{n=0}^{N-1} x[i-n] * h_\theta[n] \tag{3.2}$$

where $x[i]$ is the input signal block i and $h_\theta[n]$ a block from the impulse response. In the following, we will always use the block processing notation. This implies that the time is discretized into intervals with the same duration as a block, especially, the variable t denotes not the continuous time but the "time index" in the "unit" [block]. A more comprehensive graphical representation of the processing is depicted in Figure 3.7.

To increase the efficiency of the processing, the impulse response h is truncated to an integer multiple of the block size l so that N is also an integer. This means that we are loosing a part of the reverberant tail of the impulse response. For a typical configuration with a block size of 512 samples and a sampling frequency of 44.1 kHz, we truncate the impulse response by at most 511 samples or 11.6 ms and 4.6 %, respectively. If the position of the sound source is changing, h_θ changes instantaneously, we did not implement any fading mechanism.

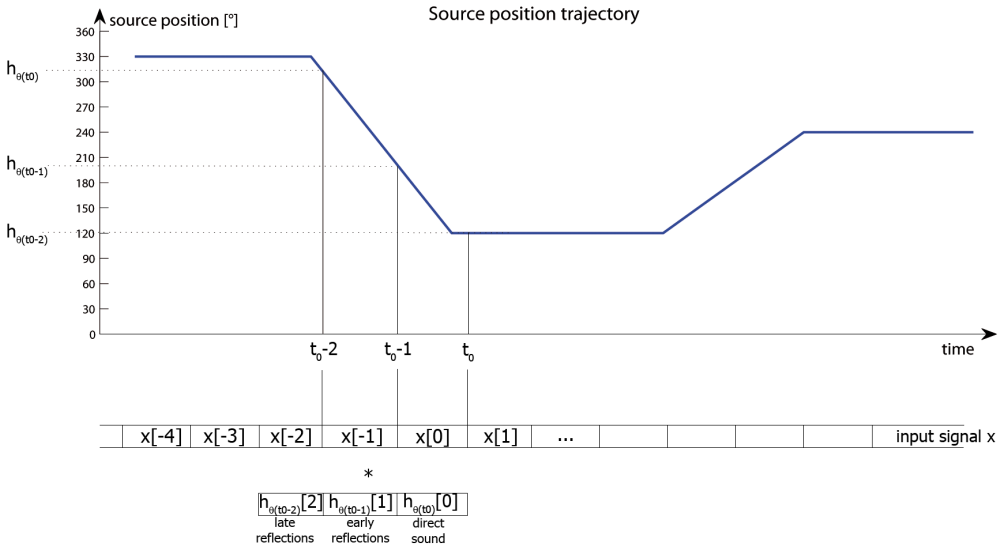


Figure 3.7: Block-wise signal processing for dynamic scenes. The impulse response $h_{\theta(t)}$ is determined by the source position θ at time t. The input signal x is processed in blocks $x[-2], x[-1], x[0], \ldots$. The impulse response h is also divided into blocks of equal length ($h[0]$, $h[1]$, $h[2]$, ...), representing different properties of the room like direct sound, early reflections, late reflections and reverberant tail. For simplicity, h has only a length of three blocks in this example. In the real system, the impulse response is much longer. The output signal block $y_{t_0}[0]$ at time t_0 is given by $y_{t_0}[0] = h_{\theta(t_0)}[0] * x[0] + h_{\theta(t_0-1)}[1] * x[-1] + h_{\theta(t_0-2)}[2] * x[-2]$. Note that every input signal block is filtered with the impulse response corresponding to the source position at the time when this very block is played.

Subjective listening tests were done with the following parameters and settings:

- Sound signals: white noise and modulated white noise
- Speed of the source: 24 °/s and 72 °/s
- Temporal resolution: 5/10/15/20/25 Hz
- Spatial resolution: 5°, 5° with linear interpolation, 1°

For a spatial resolution of 5° without any interpolation, there were audible artefacts, this resolution seems to be too coarse. With a spatial resolution of 5° and linear interpolation of the nearest neighbour impulse responses, there were some audible amplitude fluctuations in the case of a moving source. The reason for those fluctuations was that we did not took into account the different time of arrivals of the direct sound which resulted in destructive interference for some positions. For a better interpolation, we would have to use an interpolation scheme similar to the one used for HRTF interpolation described in section 2.6. Due to the real-time processing, we tried a simplified time-alignment without sub-sample precision that led to better results, but there were still audible artefacts. The more sophisticated interpolation scheme used for HRTF interpolation is computationally too expensive, therefore we decided to use pre-rendered impulse responses with a spatial resolution of 1°. A temporal resolution of at least 20 Hz was sufficient to render a smooth scene for both source speeds. Moreover, there were also no audible artefacts due to the block-wise processing, the instantaneous change of the impulse response or the HRTF interpolation.

3.2.4 Real-time Compensation of Head Movements

The next step towards a system which is able to render dynamic scenes as well as to compensate for head movements was a system which can compensate head movements but does not support moving sources. As described in section 3.2.3, the rendering of dynamic scenes with pre-computed BRIRs can be done completely offline. The extension of the system to compensate for head movements is a more challenging task since we have no prior knowledge about the head movements of a subject. We can still use pre-rendered impulse responses, but the audio signal has to be filtered in real-time. A widely used zero delay convolution algorithm which combines the efficiency of block FFT convolution with the zero delay of a direct-form processing in the time domain was considered [32]. Unfortunately, the algorithm is designed for a DSP based system and requires a real-time operating system with a scheduler that can guarantee processing deadlines. Our standard PC cannot satisfy these requirements, therefore we had to use a buffer for the audio signal and accept a small processing delay. From the rendering of dynamic scenes, we know that we need a spatial resolution of 1° (or maybe less, but we have not done any further tests) for the offline calculated impulse-responses. The block-wise processing

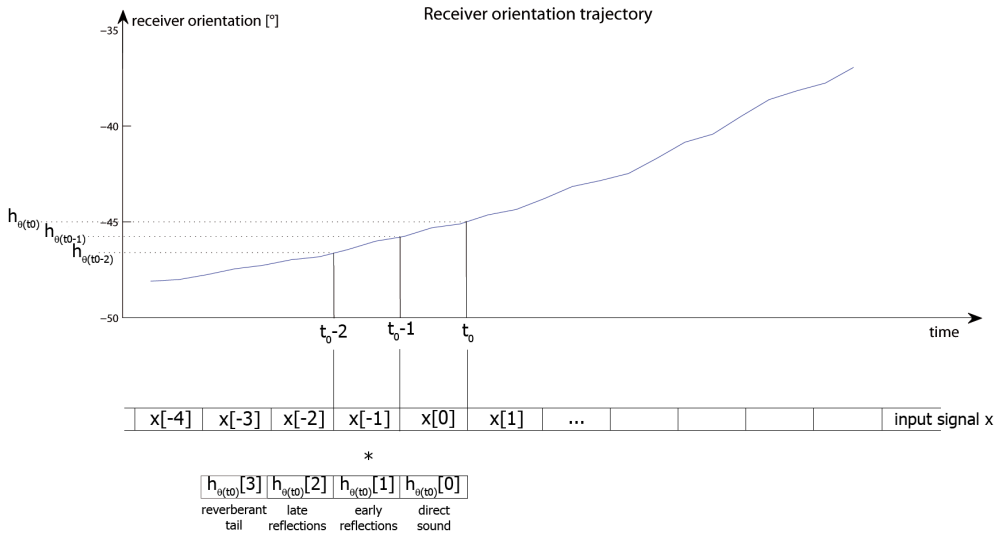


Figure 3.8: Block-wise signal processing for the compensation of head movements. The impulse response $h_{\theta(t)}$ is determined by the actual receiver orientation at time t. Ideally, the output signal block $y_{t_0}[0]$ at time t_0 is given by $y_{t_0}[0] = h_{\theta(t_0)}[0] * x[0] + h_{\theta(t_0)}[1] * x[-1] + h_{\theta(t_0)}[2] * x[-2] + h_{\theta(t_0)}[3] * x[-3]$. The difference to the processing of moving sources is that only the impulse response $h_{\theta(t_0)}$ is used instead of composing the impulse response of the blocks $h_{\theta(t_0)}[0]$, $h_{\theta(t_0-1)}[1]$, $h_{\theta(t_0-2)}[2]$ and $h_{\theta(t_0-3)}[3]$. The real processing differs from the ideal processing which is computationally too intensive. For $a = 2$, $y_{t_0}[0]$ at time t_0 is given by $y_{t_0}[0] = h_{\theta(t_0)}[0] * x[0] + h_{\theta(t_0)}[1] * x[-1] + h_{\theta(t_0-1)}[2] * x[-2] + h_{\theta(t_0-2)}[3] * x[-3]$. The consequence of this simplification is that the late reflections of the room are simulated using a somewhat outdated impulse response.

from the dynamic scenes was kept, the block size defines the size of the audio signal buffer and therefore also the processing delay and the temporal resolution (update rate). The processing delay should be as small as possible which implies a small block size and also a high temporal resolution. For every block that is played, the impulse response is updated according to the head position that is obtained by polling the head-tracker.

The differences in the signal processing compared to the simulation of dynamic scenes are limited to details. Equation 3.1 still holds, with the following differences: Obviously, the impulse responses h_θ are pre-calculated for all possible receiver orientations (head positions) instead of all possible source positions. The most important difference is that the position dependent impulse response $h_{\theta(t)}$ should now correspond to the actual receiver orientation and not to the source position at the time when the sound was emitted. Figure 3.8 shows a graphical representation of the processing.

With the available computer, a sampling rate of 44.1 kHz, a block size of 512 samples ($\approx 11.6ms$) and an impulse response length of 11025 samples ($= 250ms$), we were far away from being able to process the data in real-time. The reason is that the input signal should be convolved with the whole

impulse response every 11.6 ms, which is a too demanding task. To simplify the calculations, we filtered the input signal only with the first part of the actual impulse response which corresponds to the direct sound and the early reflections. The late reflections and the reverberant tail are filtered with a somewhat outdated impulse response. In other words, the impulse response is divided into two parts, where the first part is used for an exact processing and the second part is an approximation tho the actual impulse response. Let us assume that the first a blocks of the impulse response are used for the exact processing. The output signal is then given by

$$y[i] = \sum_{n=0}^{a-1} x[i-n] * h_{\theta(t)}[n] + \sum_{n=a}^{N-1} x[i-n] * h_{\theta(t-a+1)}[n]. \tag{3.3}$$

The first sum represents the part of the impulse response which is used for an exact simulation, the second sum represents the approximated parts. This means that the late reflections and the reverberant tail are filtered using an impulse response which corresponds to the position of the receiver (head) at the time $t - a + 1$. This introduces some sluggishness in the virtual acoustics system, listening tests will show if it's audible. Equation 3.3 does not show the simplifications that allow a faster processing. The crucial point is that in the actual implementation only the first sum has to be evaluated for every block. The second sum is a by-product of the first sum and has to be evaluated only once and not for every block.

When using this modification, a real-time processing is possible. With a block size of 512 samples (as suggested by Lentz et al. [3]) and a sampling rate of 44.1 kHz, the head position is updated at a rate of 86 Hz. The overall processing delay is then 512 samples plus another 512 samples from the soundcard buffer, in total 1024 samples or 23.2 milliseconds. The impulse response was divided into two parts as described, the boundary was set to 81.3 milliseconds (81.3 ms for precise processing, 162.5 ms for sluggish processing of the reverberant tail). In a listening test with three subjects, with a speech signal, nobody noticed any sluggishness nor artefacts even with fast head movements. As a consequence of this successful listening test, no further optimizations have been done.

3.2.5 The New Dynamic System

Up to now, we have two working schemes, one for the rendering of dynamic scenes (moving sources) and one for the compensation of head movements. The goal is now to combine the two to get a system which is able to do both tasks.

Both systems have only one degree of freedom, with a spatial resolution of 1°, we have to calculate and store 360 binaural impulse responses. The simple combination of both systems leads to a system with two degrees of freedom. This would require to store 360^2 impulse responses. With an impulse response length of 11025 samples, two channels and 64 bits per sample, this would require a memory capacity of $360 \cdot 360 \cdot 11025 \cdot 8 \cdot 64$ bits or 21.3 GB, which is far too much. To reduce the amount of required memory, we used the fact that our

test room (cf. section 3.1.4) is almost symmetric. A source movement can be modeled by a head movement in the opposite direction and vice versa. In a highly asymmetric room, this would probably lead to an inaccurate simulation, to what extent it would be audible is an open question.

The processing of a moving source and a head movement differs mainly in one detail: For a source movement, the input signal block is filtered with the impulse response corresponding to the source position *at the time when this block is played*. This results in an impulse response which is composed of several blocks of different impulse responses (corresponding to different source positions). The resulting impulse response is therefore not only dependent on the actual source position but also from former source positions. The reason is that it takes some time for the sound to travel trough the space - late reflections which originate from a source at position 1 have to be filtered with the impulse response corresponding to that position, at the same time the direct sound from the source, which has already moved to position 2, has to be filtered with the impulse response that belongs to position 2. In the case of a head movement, the impulse response is only dependent from the actual head position. The fact that a) the reverberant tail of the room impulse response dos not change significantly for small source movements, b) the late reflections and the reverberant tail are rendered with a somewhat outdated impulse response in the case of a head movement anyway and c) our system is running at a high update rate of 86 Hz led to the decision to not take into account this small difference and simply use the processing scheme for the head movement compensation. Furthermore, the human auditory perception is inherently sluggish, the faster a source moves, the less accurate the localization is [29]. This means that for a fast head and/or source movement, where the error of the simplified processing becomes larger, the auditory system is not very accurate. For slow movements, the error is also small because the impulse responses from neighbouring positions are very similar.

This system was used to perform all the listening tests described in chapter 4.

3.3 Limitations

The system has several limitations, namely

- The open ITE micro speakers have a limited frequency range, they cannot emanate frequencies below $\sim$ 350 Hz. All signals that were played over the loudspeakers as well as with the simulation are therefore bandpass filtered in order to allow a fair comparison between the simulation and real loudspeaker sources.

- The system can only compensate for head movements in the horizontal plane. Source movements were also restricted not only to the horizontal plane, they are restricted to a circle around the test subject.

- The system can only simulate dry rooms with a short reverberation time,

a longer reverberation time results in a longer room impulse response which increases the computational effort.

- For a highly asymmetric room, the system would probably not be able to produce satisfactory results.

All limitations except the limited frequency range are a consequence of the economical, flexible Matlab-based software solution. There is always a trade-off between performance and flexibility. A system with dedicated hardware and heavily optimized software could not only perform a real-time filtering of the audio signal but also a real-time computation of the binaural room impulse response and would therefore not be affected by the mentioned limitations.

3.4 Practical Issues

Although this report is in no way a manual for the virtual acoustics software, some important practical issues are described here.

- For an efficient calculation of the convolution, the Matlab function fftfilt is used. This is a highly optimized function that performs the filtering in the frequency domain based on the fftw library [33]. The difference to the convolution operation – apart from being much faster – is that the result of `conv(A,B)` is of `LENGTH(A)+LENGTH(B)-1`, while the result of `fftfilt(A,B)` is of `LENGTH(B)`. Omitting the details of the implementation, this has the same effect in the output signal as if the `conv` function would be used with an impulse response that is truncated by one block. An impulse response of length 11025 samples in combination with a block size of 512 samples is first truncated to an integer multiple of the block size, in that case to 10752 samples. Then, it is further "truncated" by one block to the final "usable" length of 10240 samples. As mentioned, not the impulse response is truncated, but the outcome of `fftfilt(x,`$\mathtt{h}_{\mathrm{full}}$`)` is the same as `conv(x,`$\mathtt{h}_{\mathrm{truncated}}$`)`.

- The head-tracker crashes regularly after initialization, leading to wrong measurements. The software can detect this because the sequence of the crashes is always following a certain scheme. First, the head-tracker acts as it should. Then, the output looks like white noise with a high amplitude. After that phase, the output of the sensor is almost zero. After each initialization of the head-tracker, five consecutive values are queried. If the sum of the differences between the consecutive values is below a certain threshold (near-zero output below the internal noise) or exceeds another threshold (noisy output with a too high amplitude), then the operation is stopped. The sensor has then to be removed from the computer and plugged in again – it is then usually working again for a certain number of initializations.

Chapter 4

Evaluation

The goal of the tests described in this chapter was to evaluate the system, that is, to make sure that the system is working as it should and could be used for the evaluation of hearing aid algorithms. This implies that we did not focused the tests on auditory skills of the test subjects (like the ability to localize sound sources), but to test if the generated scenes are perceptually convincing, free of processing artefacts or distortions in the case of head-movements and/or dynamic scenes. We set up two tests, one with static scenes and one with dynamic scenes. For both tests, we presented sound over the loudspeaker and over the virtual acoustics system and compared the results.

During the implementation of the system, it turned out that the motion tracking sensor is not as accurate as one could expect. Hence, this chapter starts with a section about the accuracy of the head-tracker.

4.1 Accuracy of the Motion Tracking Sensor

4.1.1 Test Setup and Results

The compensation of head movements requires an accurate, drift-free motion tracking sensor. The specifications of the sensor that was used in this work are described in chapter 3.1.1. During the development of the system, we noticed that the sensor is far away from being drift-free, in contrast to the claims of the manufacturer. A quick estimate of this drift was done in a two-step procedure:

1. Movement phase of 30 seconds where the sensor was moved by hand.
2. Measurement phase where the sensor was kept still and not moved anymore for 15 seconds.

In the movement phase, the sensor was moved by hand along the yaw axis, i.e., it was turned around the axis which is used for the measurement of head movements in the horizontal plane. This axis is the most important axis and also the only one which suffers from the drift. The sensor was moved for 30 seconds with a constant angular speed. After 180° degrees, the direction of rotation was changed, the resulting trajectory is a triangular wave, an example is depicted Figure 4.1. In step two, right after the movement phase, the sensor was put

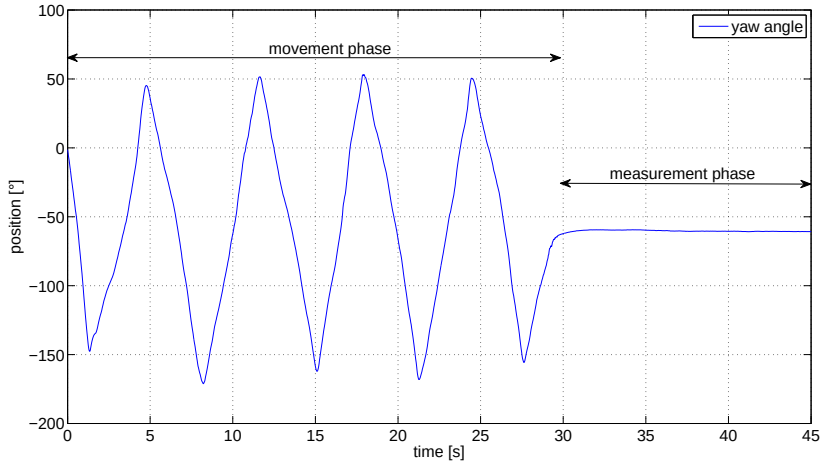


Figure 4.1: A typical, hand-made trajectory of the sensor drift test, recorded with the drift-free MTi sensor.

on a table for 15 seconds and the position (which should remain constant) was measured. The movement of the sensor by hand is not very accurate and the movement scheme is not a natural movement, nevertheless it allows a coarse estimate of the sensor drift. Each measurement was repeated three times and the average RMS error was calculated for the time when the sensor laid on the table. The RMS is defined as

$$RMS = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y[k] - y[0])^2} \tag{4.1}$$

where n is the number of samples read from the sensor and $y[k]$ the yaw angle.

We had the possibility to borrow an Xsens MTi 3DOF Orientation Tracker[1] and to do the same test with this sensor. The MTi sensor is fully compatible with the MTx sensor used in this work, but it is additionally equipped with a gyroscope. Table 4.1 lists all settings of the short sensor test, cf. also section 3.1.1. Table 4.2 and 4.3 list the results and Figure 4.2 shows two typical curves of the MTx sensor during the measurement phase, when the sensor was not moved.

[1]The sensor was provided by Bernd Tessendorf from the Wearable Computing Group at ETH Zurich. This group has a long lasting experience with motion capturing sensors and they use the MTi sensors as a reference when testing other sensors.

Sensor settings	
Filter Gain	1.0
Weighting factor	1
Adapt to Magnetic Disturbances	off
Sensor sample frequency	100 Hz
Test procedure	
Moving period	30 s
Measurement period	15 s
Number of repetitions	3
Angular velocity 1	20 °/s
Angular velocity 2	45 °/s
Angular velocity 3	90 °/s
Angular velocity 4	180 °/s

Table 4.1: Parameters and settings used for sensor drift test

angular velocity	run 1	run 2	run 3	average RMS error
20 deg/s	4.92	7.03	8.01	6.65
45 deg/s	14.07	7.24	23.56	14.96
90 deg/s	24.92	28.76	24.33	26.00
180 deg/s	26.01	29.59	27.63	27.74

Table 4.2: RMS errors in [°] of the **MTx** sensor in phase 2 of the drift test

angular velocity	run 1	run 2	run 3	average RMS error
20 deg/s	0.45	0.28	0.14	0.29
45 deg/s	0.17	6.73	5.93	4.27
90 deg/s	1.26	1.05	1.12	1.14
180 deg/s	1.13	2.30	2.57	2.00

Table 4.3: RMS errors in [°] of the **MTi** sensor in phase 2 of the drift test

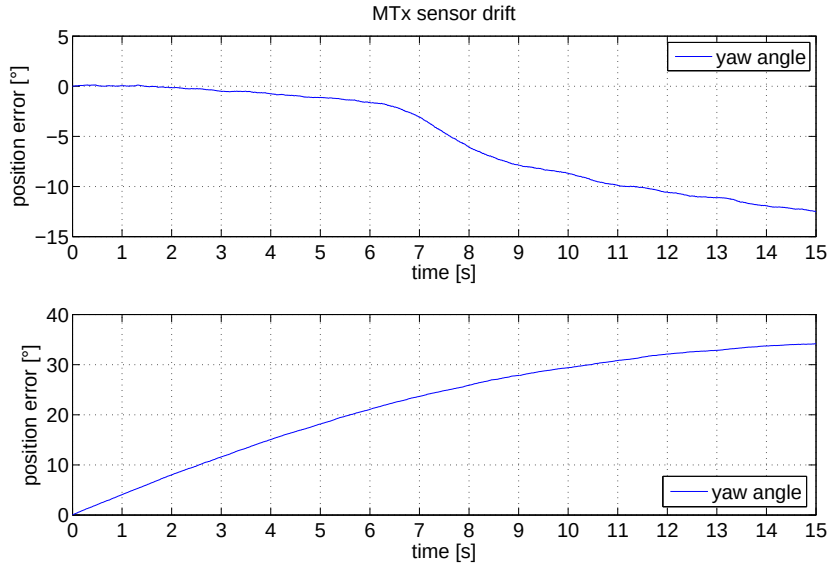


Figure 4.2: Example of MTx sensor drift. The angular velocity in the moving phase was 20 deg/s for the top curve and 90 deg/s for the bottom curve.

4.1.2 Implications

The results indicate that the Xsens MTx sensor suffers heavily from drifts, which can result in a deviation in the horizontal plane of 30° and more after 15 seconds. The yaw axis was the only axis that showed such a behaviour. If the sensor was orientated in a different way, the drift has not disappeared. Different settings for the sensor fusion algorithm did not help either. The test was repeated in a different room with similar outcomings. A thorough test of the performance of a Xsens Moven Motion Capturing Suit that is also equipped with MTx sensors was done by Damgrave and Lutters [34]. They conclude that one of the major problems of the MTx inertial motion capturing sensor is the drift on the horizontal plane, which is consistent with our findings. Their explanation of this phenomenon is that the earth magnetic field which is used by the sensor to determine the horizontal position is too weak.

The dynamic error of the sensor, i.e., the error during a movement, was not investigated due to the absence of any suitable test equipment and because it would have been too far away from the scope of this work. The listening tests described in the next sections were done with the inaccurate MTx sensor. In preliminary listening tests, the performance and accuracy of the MTx sensor in the virtual acoustics system with natural head movements was considered as good enough to proceed with the listening tests. However, we would strongly recommend to replace the MTx sensor with a MTi sensor. The MTi is a gyro-enhanced version of the MTx sensor which is fully compatible and does not show any drift.

4.2 Subjective Listening Test with Static Scenes

For a first evaluation, we decided to use static scenes. The two main reasons for this decision were:

1. Separate the evaluation of the head-movement compensation and dynamic scenes.

2. It is much easier to assess static scenes, because there are no dynamic psychoacoustic effects present (cf. section 2.4.1).

From previous listening tests with the virtual acoustics system, we knew that the externalization of sound sources is often not fully convincing, i.e., that sound sources are perceived in the head. Another outcoming of these tests was an increased rate of front-back confusions compared to scenes presented over loudspeakers [1]. These two issues motivated - among other reasons – the development of the extensions described in this work. Therefore, one goal of this listening test was to assess the externalization when listeners move their head. To test if the motion tracking sensor and the signal processing is fast enough to render a convincing scene where listeners cannot hear any sluggishness when they move their head, they had to assess the stability of the sound source. Finally, the test subjects had to decide whether the source was simulated or played over the loudspeaker. The more confusions they make, the better the virtual acoustics system is.

4.2.1 Test Subjects

A total of six individuals served as volunteers for this listening test. There were four normal hearing male subjects, one normal hearing female subject and one hearing impaired male subject. The hearing was verified by standard clinical audiometry. All of them except the hearing impaired participated earlier in a localization experiment which was based on the same hardware and room simulation software (but without the head-movement compensation), so they are experienced listeners. The were aged 25-48 years with a mean of 35 years.

Three additional test subjects, two male and one female, participated in this test, but they did not made a retest. One of them had a mild hearing loss. They were also experienced listeners. Because of the high test-retest reliability, their results are included anyway. They were aged 32-42 years with a mean of 37 years.

Figure 4.3 shows the audiograms of the two hearing impaired test subjects.

4.2.2 Stimuli

For this test, a speech signal and a noise signal was chosen as sound stimulus. The speech signal was chosen to represent a natural signal which is easy to localize and allows a meaningful rating of the externalization. Simulated

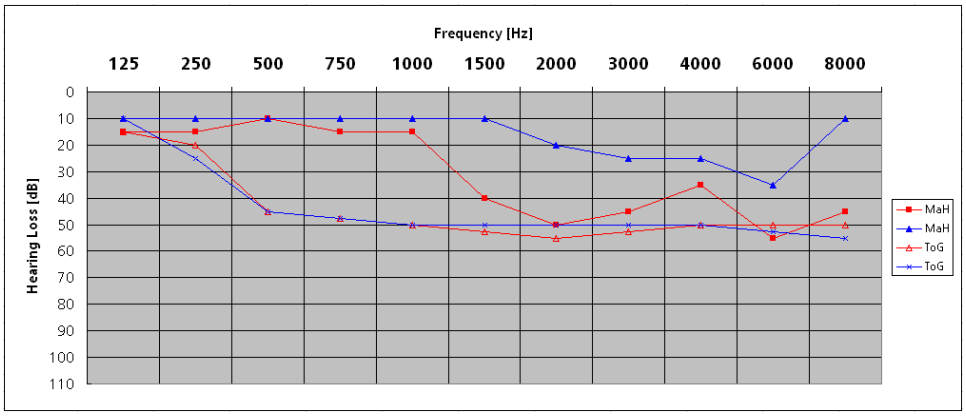


Figure 4.3: Audiograms of the two hearing impaired subjects

sources have a slightly different coloration and also not exact the same loudness as real loudspeaker sources, therefore the amplitude and coloration of the various stimuli were randomly changed. These two cues which easily allow the discrimination of the simulation from a real source could therefore not be used by listeners. We decided to take a speech signal spoken from different male speakers. It was taken from the Timit database [35] and consisted of two sentences, the duration of both sentences varied from 5 to 8 seconds, depending on the speaker. The noise was used as a test signal because a noise signal is of a constant amplitude and contains no fluctuations like the speech signal. It is therefore better suited to detect some processing artefacts like amplitude variations or other distortions. The noise was randomly colored. The sounds were presented at 60 $\pm$1 dB SPL for the normal hearing subjects and the one with the mild hearing loss. The sounds for the hearing impaired subject with the moderate hearing loss were played at a level of 70 $\pm$1 dB SPL. All sounds were bandpass-filtered to remove the frequencies which cannot be produced by the open ITE speakers. The lower cutoff frequency and the higher cutoff frequency, respectively, was 400 Hz and 8000 Hz, respectively.

4.2.3 Parameters

The parameters for the dynamic virtual reproduction system were set according to table 4.4

For the convenience of the reader, a short summary of these parameters is given here: The sampling frequency is the sampling frequency of the input sound file and the output sound which is sent to the soundcard. The lower and upper cutoff frequencies are not parameters, but are the result from a band-pass filter applied to the input sound file. The reason for this band-pass filter is to remove frequencies which cannot produced by the open ITE speakers. The block size refers to the block-wise processing described in section 3.2.3. The soundcard buffer size is the sound card's internal buffer size. The overall processing delay is the delay between a position update from the headtracker until the sound filtered with the impulse response belonging to that position

Parameter	**value**
sampling frequency	44.1 kHz
lower cutoff frequency	400 Hz
upper cutoff frequency	8000 Hz
block size	512 samples
	11.6 ms
soundcard buffer size	512 samples
	11.6 ms
overall delay	1024 samples
	23.2 ms
effective room impulse response length	10240 samples
	232.2 ms
early reflection part	3584 samples
	81.3 ms
head tracker sampling frequency	100 Hz
head position update rate	86 Hz

Table 4.4: Software parameters used in static listening test

is actually played. The effective room impulse response length is the "usable" room impulse response filter length. The sound is filtered exactly only with the first part of the actual impulse response, denoted as early reflection part. For the reverberant tail, an outdated impulse response is used to filter the sound. The head tracker sampling frequency is the internal sampling frequency of the motion tracking sensor. The head position update rate is a redundant information since it is the inverse of the block size.

4.2.4 Procedure

The test was divided into 2 rounds of 16 trials plus one single test trial in the beginning. The purpose of the test trial was to familiarize the subject with the test procedure and to make sure that he has understood the test procedure. In the first round, 16 trials with speech signals were presented randomly over loudspeaker or with the simulation, under the following conditions:

- 2 types of sound sources: simulated and real sources
- 4 Positions (front/right/back/left)
- each position was served twice
- 8 different speakers. Each speaker was used once for a loudspeaker presentation and once for a simulation presentation.

which results in a total of 16 trials. In the second round, noise signals were used. The sound was repeated until the subject interrupted the sound output. The subject was encouraged to move his head even if he could easily localize

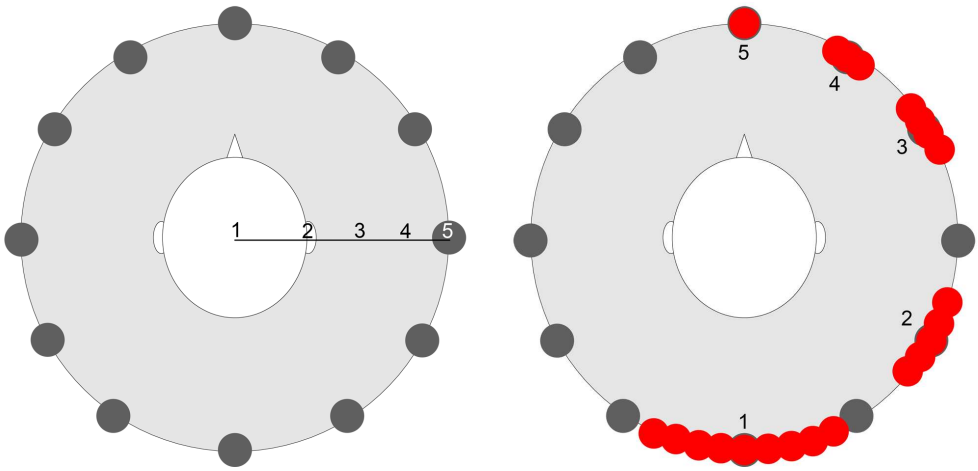


Figure 4.4: Answer maps for questions about externalization and stability

the sound source, but the head-movements were not compulsory. The task of the subjects was to answer the following questions:

1. On a scale from 1-5, do you hear the sound source in your head or from the loudspeaker?
2. On a scale from 1-5, does the sound source remain stable if you turn your head?
3. Where does the sound come from: Loudspeaker or simulation?
4. (If the answer to question 3 was "simulation"): Why? Any further remarks?

The scales of the question 1 and 2 were illustrated with the Figures 4.4.

The instructions were given in German, the original instructions can be found in Appendix B. During the tests, it turned out that the instructions are hard to understand because they contain too much information. There were no control questions to ensure that all test subjects fully understand their task. A proposal for improved test instructions can be found in appendix B.2. The whole test took about 20-30 minutes, depending on how long the subjects listened to the sounds.

A retest was done after 1-2 months.

4.2.5 Results

Externalization

The ratings of the externality are shown in Figure 4.5 and Figure 4.6. Figure 4.5 shows the results of all subjects, test and retest together. Figure 4.6 shows the results of the test and the retest separately, based on those six subjects who did the retest. The corresponding numerical values are listed in Table 4.5.

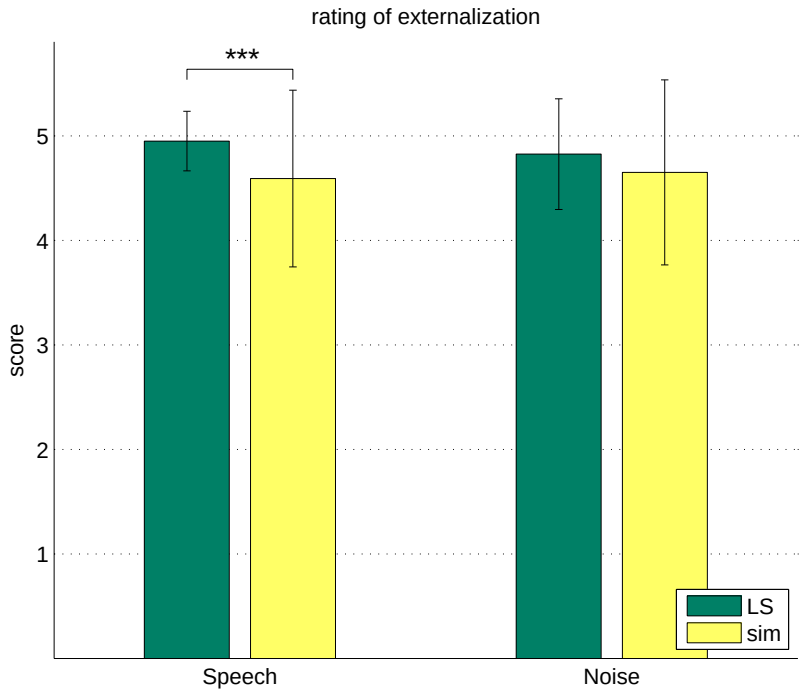


Figure 4.5: Rating of externalization for all subjects

Significance was tested with the Wilcoxon matched pairs signed rank test. The details of this statistical test are explained in Appendix C.3.

During the tests, it turned out that sound sources often appeared in the head, but as soon as the subjects moved their head, the source "leaved" the head and was well externalized (and remained external even if the subject returned to the starting position). Test subjects asked if they should rate the externalization of their first impression or the externalization after the head movement. They were told to rate the externalization during the head-moving phase. This instruction probably improved the mean externalization rating. However, a source which appeared in the head, even if only in the first few seconds of a sound presentation, was often used as a cue to detect the simula-

		Test	Retest	Avg
Speech	LS	4.9	5	4.95
	sim	4.44	4.69	4.59
Noise	LS	4.83	4.94	4.83
	sim	4.69	4.56	4.65

Table 4.5: Rating of externalization. Test and retest values are based on 6 subjects, the average includes also the results from the 3 subjects who have not done a retest

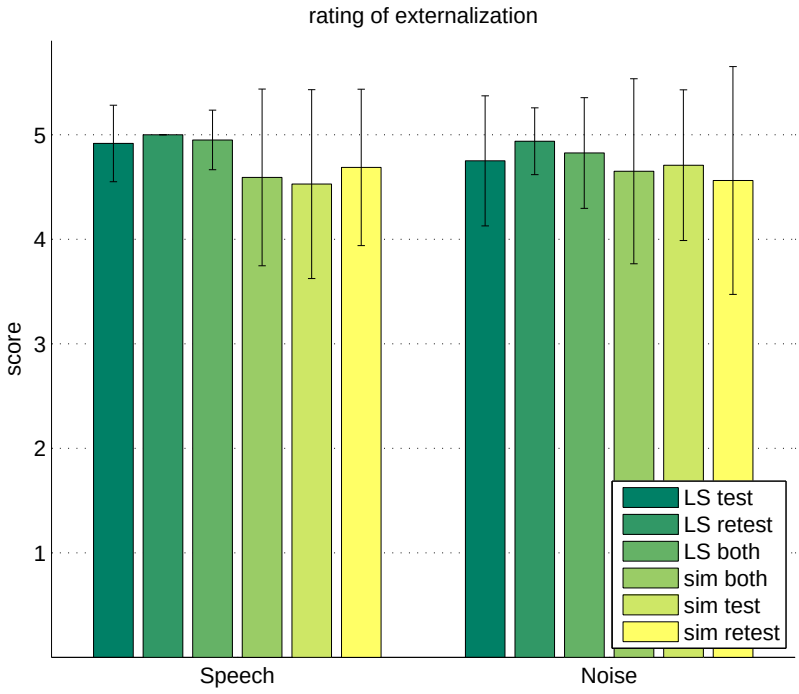


Figure 4.6: Rating of externalization for 6 subjects, test and retest separately. No statistical analysis was performed: The center bars correspond to Figure 4.5 and the test-retest relibaility was evaluated separately.

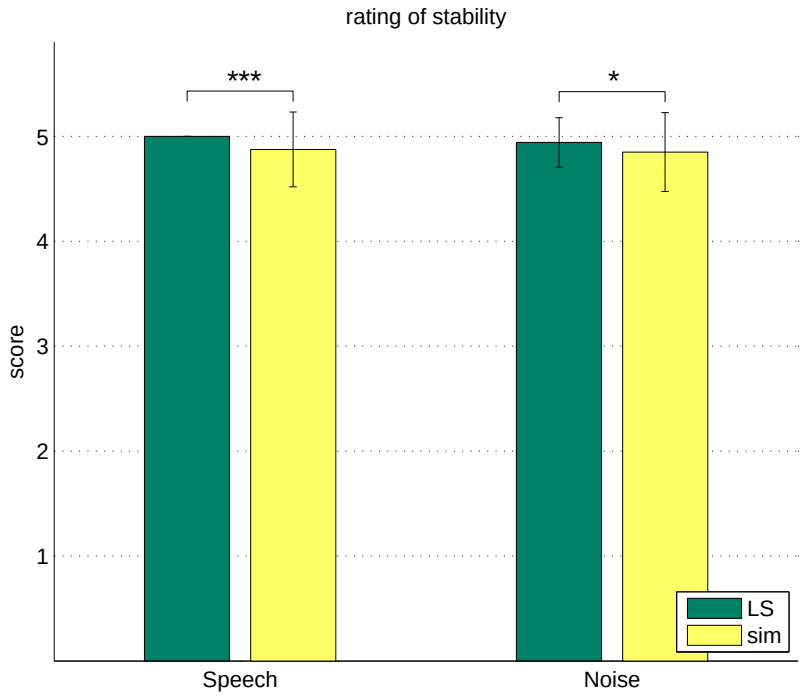


Figure 4.7: Rating of stability for all subjects

tion. This can also be seen in the results of question 4.

Stability

The ratings of the stability are shown in Figure 4.5 and Figure 4.6. Figure 4.5 shows the results of all subjects, test and retest together. Figure 4.6 shows the results from the test and the retest separately, based on those six subjects who did the retest. The corresponding numerical values are listed in Table 4.6. Significance was tested with the Wilcoxon test.

		Test	Retest	Avg
Speech	LS	5	5	5
	sim	4.88	4.88	4.88
Noise	LS	4.98	4.98	4.94
	sim	4.92	4.81	4.85

Table 4.6: Rating of stability. Test and retest values are based on 6 subjects, the average includes also the results from the 3 subjects who have not done retest

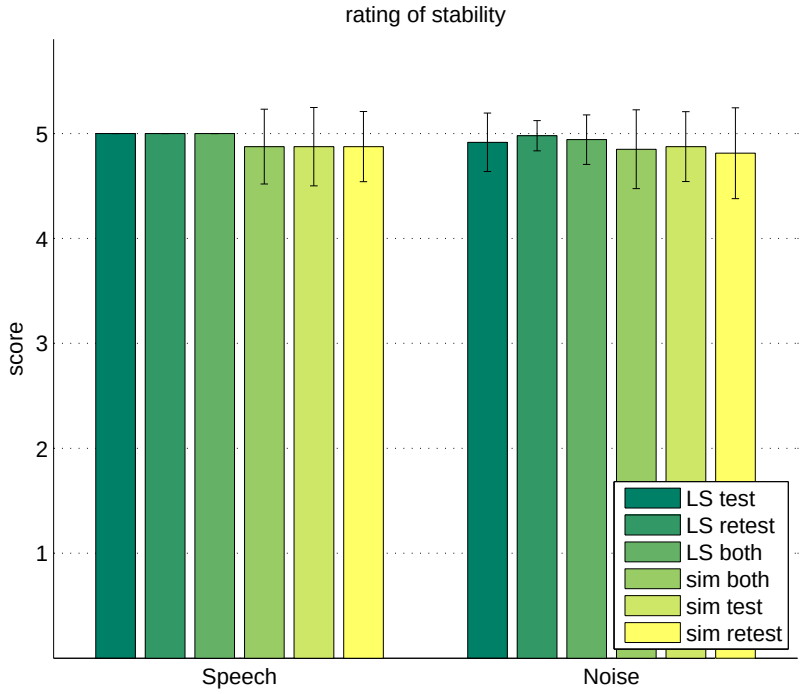


Figure 4.8: Rating of stability for 6 subjects, test and retest separately. No statistical analysis was performed: The center bars correspond to Figure 4.7 and the test-retest relibaility was evaluated separately.

Speech		playback over	
		LS	sim
classified as	LS	111	43
	sim	9	77

Noise		playback over	
		LS	sim
classified as	LS	100	69
	sim	20	51

Table 4.7: Confusion matrix

Classification of the Sound Source

The results from question 3 are shown as a matrix of confusion in Table 4.7. The results from the open question, where the subjects should reason their decision in question 3 (only if the answer was "simulation") are summarized in Figure 4.9

The reasons for the decision "simulation", as reported by the test subjects, can be combined into the following groups:

- **Poor Externalization** means that the sound source was not perceived external. A variant which often occurred was that only the first impression of a sound source was perceived in the head, but as soon as the subject moved his head, the source was perceived as external.
- **Stability** means that the source remains not perfectly stable if the subject turns his head. This is the result from the dynamic sensor error.
- **Off-Position:** An off-position sound source is a source which is located somewhere between two loudspeakers, which has to be a simulated source. A static sensor error results in an off-position source. This is probably the consequence from the sensor drift (cf. section 4.1).
- **Artefacts:** Audible processing errors, resulting in a stuttering signal or in very short pauses (dropouts).
- **Unnatural Frequency Response:** Unnatural frequency response, e.g. too few low frequencies or an unnatural change in the frequency response during head movements.
- **Diffuseness:** An imprecise, diffuse sound source, a sound source which seems to be not a point source but rather a source that is spread over an unnatural large area.
- **Front-back Uncertainty:** Some listeners were not sure if the source was in the front or in the back (front-back confusion) and concluded that such an uncertainty stems from the simulation.

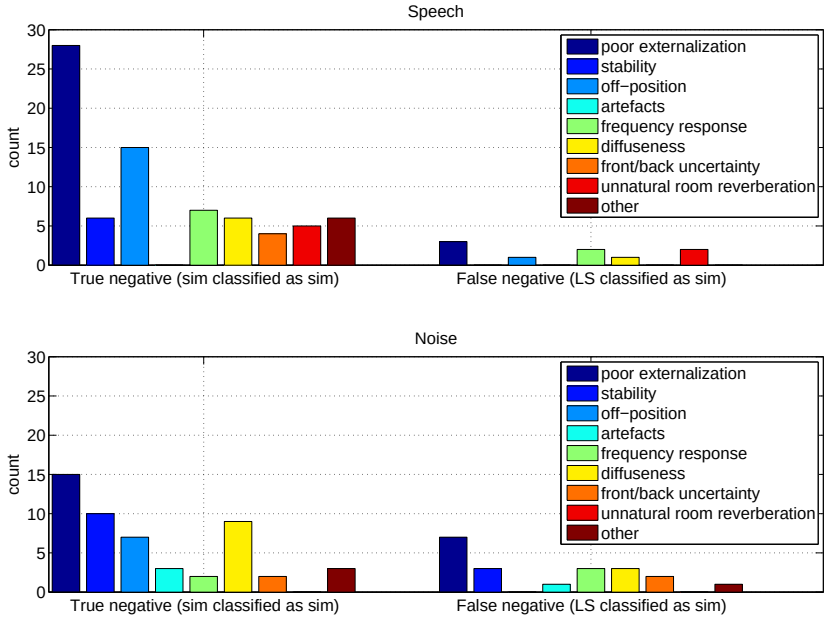


Figure 4.9: Reasons for classification as simulation

- **Unnatural Room Reverberation:** One test subject reported that the room sounds too dry with the simulation.

Position Dependency

Apart from the dependency of the stimuli, the results depend also on the playback position. The externalization ratings in dependence of the source position are depicted in Figure 4.10. Table 4.8 shows the confusion matrix with percentage values instead of absolute values, because only a quarter of all presentations came from the front position. Finally, Figure 4.11 shows the reasons for the classifications as simulation.

The results of the positions at 90°/180°/270° azimuth do not show any significant differences, therefore they were combined. The stability rating is also not dependent on the playback position.

Test-Retest Reliability

6 of 9 subjects did a retest, on average 54 days after the test. To determine the test-retest reliability, the percentage of coincident ratings in the test and in the retest was calculated for every subject and for both ratings. The results are listed in Table 4.9

The values of 0.85 and 0.94 can be considered as a good test-retest reliability and therefore, the results from those subjects who did only one test are also included in the sections above.

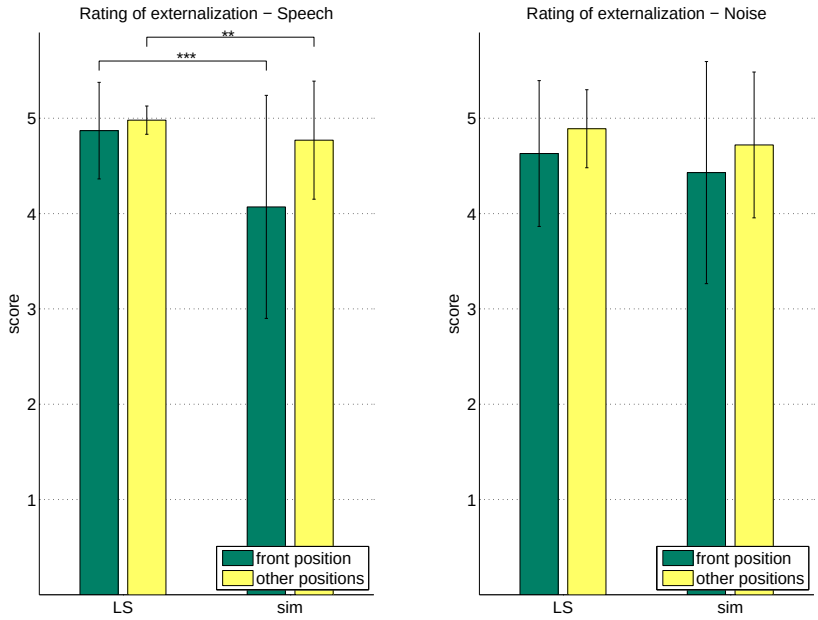


Figure 4.10: Rating of externality for the front position and the other positions. Results from all subjects.

Speech		LS		simulation	
		front	other	front	other
classified as	LS	87	94	20	41
	sim	13	6	80	59

Noise		LS		simulation	
		front	other	front	other
classified as	LS	83	83	60	57
	sim	17	17	40	43

Table 4.8: Confusion matrix for the front position and the other positions, values in [%]. Results from all subjects.

	Externalization			Stability		
	LS	sim	avg.	LS	sim	avg.
Speech	93.8	70.8	82.0	100	87.5	94.0
Noise	93.8	83.3	89.0	100	87.5	94.0
Average	85.0			94.0		

Table 4.9: Test-retest reliability in [%]

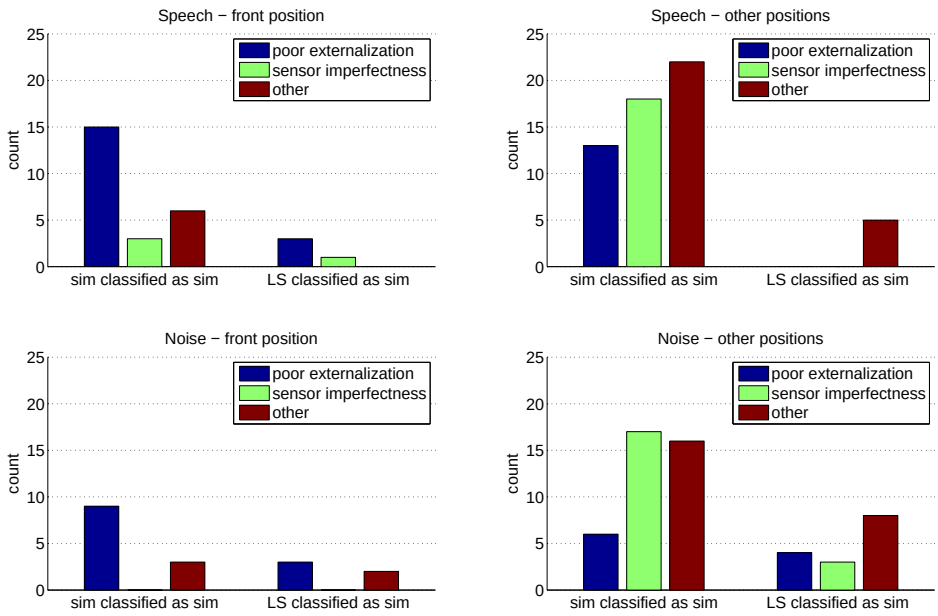


Figure 4.11: Reasons for classification as simulation in dependence of the playback position. Note that there are 30 front position presentations and 90 presentations from the other positions. The reasons are grouped, the "sensor imperfectness" category includes unstable and off-position sources, all the other reasons (except "poor externalization") are subsumed in the "other" category.

Learning Effect

Although all test subjects had previous experiences with the virtual acoustics system, we wanted listeners that are as naive as possible. The absence of a training phase with feedback was a result of this demand. One test subject mentioned a learning effect during the test. He claimed that he could learn to distinguish between simulation and loudspeaker, although no feedback was given during the experiment. To determine if there is any learning effect, we analyzed the evolution of the number of classification errors during a test session. A classification error is a loudspeaker source that was classified as simulation or a simulated source that was classified as a loudspeaker source. The off-diagonal elements in the confusion matrix (cf. section 4.2.5) are another representation of these false classifications. If there are significant less classification errors in the end of a test session than in the beginning, then there is a learning effect.

We used linear regression analysis to find out if there is such an effect. Linear regression is a least squares estimate of a linear regression model, that is, in our case, a linear relationship between the trial number and the number of false classifications. Furthermore, the regression analysis gives an evidence about the confidence of the estimate. A formal description of the linear regression analysis is given in appendix C.1. We make two assumptions about the learning effect:

1. The learning effect is linear, that is, the number of classification errors decreases linearly over time. For the short duration of this test, this assumption might be reasonable, but for a longer test, the learning curve has perhaps the form of an exponential decay. By taking the logarithm of all values in the relevant equations, we can apply the linear regression analysis again to detect if there is an exponential decay in the number of classification errors over time.

2. The learning effect (if there there is one) for the two stimuli, speech and noise, is independent. If a listener has learnt to detect the simulation of a speech signal, he is still not able to detect the simulation in the case of a noise signal.

For the assumption of a linear learning curve, we are looking for a line

$$y = bx + a.$$

For the assumption of an exponential learning curve, the regression line is of the form

$$y = ae^{bt}$$

which is equivalent to the linear equation

$$ln(y) = ln(a) + bt.$$

The resulting coefficients from the regression analysis are listed in Table 4.10, together with the values for R^2 and the significance level. R^2 can take values in the range $[0, 1]$ and is an estimate of the "goodness of fit" of

	Speech				Noise			
model	a	b	R^2	sig.	a	b	R^2	sig.
linear	4.775	-0.179	0.199	0.083	6.625	-0.125	0.076	0.302
exponential	3.898	-0.038	0.113	0.204	6.510	-0.028	0.095	0.246

Table 4.10: Results of the regression analysis

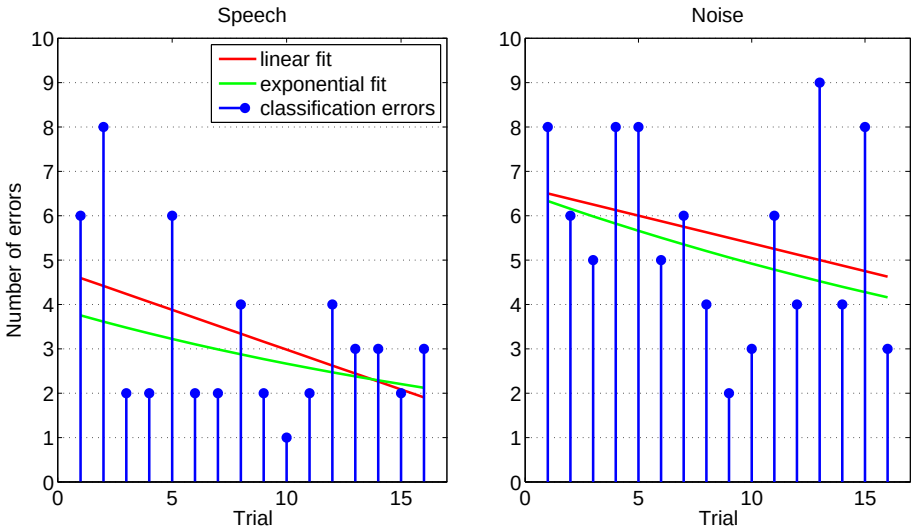


Figure 4.12: Graphical representation of the regression analysis

the line. It represents the percentage of the variation of the data explained by the fitted line; the closer the points to the line, the better the fit. A value of 1 corresponds to a perfect fit, that is, all data is on the line. The significance level equals the probability that the observed learning effect and the resulting line, respectively can be explained by coincidence. For a value smaller than 0.05 it is generally accepted that the outcome of the regression analysis is statistically significant. Figure 4.12 visualizes the regression lines and the evolution of the classification errors during the listening test (for all subjects, test and retest together).

There is no significant learning effect in any case, but the quality of the regression is rather poor with low values for R^2 due to the large variance in the data. It can be concluded that there is not enough data to quantify a learning effect.

4.2.6 Discussion

The results show that the dynamic virtual acoustics system performs quite well under static conditions.

Considering the large sensor drift (cf. section 4.1) the whole system performs surprisingly good. Approximately in one third of all correct simulation detections, the stated reason was an unstable source or an off-position source.

The human auditory system can not detect small movements of a source, the minimum audible movement angle is larger than 8° even for sources located at the most sensitive front position (cf. section 2.4.2). This would explain why the sensor drift has only a moderate influence on the overall system performance. It is also not known how large the sensor drift for real head movements is – and these real movements differ from person to person. Although the sensor inaccuracy does not result in a total unusable system, a more accurate sensor would further improve the performance of the system.

Another issue is the externalization. The simulated sources are not always fully external. To be precise, in most presentations, the source is either in the head or well externalized – there are almost no shades of grey. Especially for the front position, the externalization is often not convincing, at least in the case of a speech signal. For noise signals, there are no significant differences between real and simulated sources. The reason for this outcoming is that the human auditory system is most sensitive for to sources located in the front – and even a very small imperfectness of the HRTF for this position can lead to a in-head impression. Subjects often reported that they hear the source in their head until they turn their head, and that with head movements, the source "leaves" the head. Even if they returned to the initial position, the externalization often persisted.

There is a weak, but not significant evidence that one could learn to distinguish between the simulation and real sources. Additionally, the two most experienced listeners (the two main developers of the system) made on average 1.5 false classifications with speech signals and 4.5 false classifications with noise signals, respectively, whereas the other subjects made 4.2 and 6.5 false classifications, respectively. This supports the hypothesis that there is a learning effect. Due to the small number of test subjects, the results are affected by these two expert listeners.

The influence of a hearing loss was not investigated since only two hearing impaired subjects participated in the test. Furthermore, the results do not depend on the stimulus (the speaker and the noise coloration, respectively).

In summary, the head movments resolve front-back confusions, sound very natural and improve the externalization compared to the old system. With a better sensor, the performance of the system could be further improved.

4.3 Listening Test with Moving Sources

One goal of this work was to make it possible to render dynamic scenes (i.e., moving sources), with the intention to investigate localization under realistic, dynamic conditions. The following test should provide some informations if the extended system is suitable to conduct such dynamic localization experiments. The general idea behind this test was the same as for the static test: Compare some performance measures of simulated sources with real loudspeaker sources.

4.3.1 Test Subjects

The test subjects were the same as for the static listening test, but without the three one-time participants, so there were six individuals, four normal hearing male, one normal hearing female and one hearing impaired male with a moderate hearing loss. They all did a test and a retest, on average 37 days after the test.

4.3.2 Stimuli

It would be obvious to use the same stimuli as in the static test to allow a comparison with these results. One problem that arises when presenting dynamic scenes is that a subject has no possibility to detect a change in the direction or velocity of a sound source when there is no sound signal present. A speech signal usually contains short pauses of up to 0.5 seconds, which led us to the decision to use only continuous noise signals. The noise was randomly colored and presented at 60 $\pm$1 dB SPL for the normal hearing subjects and 70 $\pm$1 dB SPL for the hearing impaired subject. It was bandpass filtered to remove the frequencies which cannot be produced by the ITE speakers. The frequency range was 400 Hz to 8000 Hz. The software parameters were the same as in the static listening test, they are listed in Table 4.4.

4.3.3 Procedure

There is no single test for the localization of moving sounds, but the test design highly depends on the subject of interest. There are three main fields of interest described in literature (cf. also section 2.4): Many studies on questions of motion detection and discrimination, also known as minimum audible moving angle, were conducted. In these studies, subjects have typically been asked to discriminate between directions of motion or to discriminate between a stationary and a moving sound source. A second topic that recently became more popular, is the perception of moving sound. The Fröhlich effect and the representational momentum are two effects of mislocalization of moving sounds (and moving visual stimuli, too). In a typical experiment setup, listeners had to make a relative judgment, that is, if the sound source is left or right of some reference point. Another possibility is to let the subjects align a hand pointer in the direction of the sound source. A third type of experiment deals with the contribution of head motion cues to localization. Experiment setups include classic localization (i.e., identifying one of several numbered sources) [14], verbal expression of apparent azimuth and elevation [2], forced choice procedures for short stimuli and a setup which is called "pointing with the nose" [38]. There is an open-loop version of this task where the stimuli are so short that listeners start to move their head after the stimulus ceased. In the closed-loop version, longer stimuli are used such that listeners have enough time to face the source.

Our goal was to test the head-tracker functionality in combination with moving sources, that is, head movements *and* source movements. All of the above mentioned tests were either used to investigate the localization of static

sources in dependence of different types of head movements or to explore the localization of dynamic sources without head movements. We decided to use the closed-loop version of "pointing with the nose" together with moving sources to test the dynamic virtual acoustics system. The head-tracker was used not only to compensate for but also to record the head movements of a test subject. For a comparison of the simulation with loudspeaker presentations, we had to render moving sources also with our loudspeaker array. We used vector base amplitude panning (cf. section 2.7) to generate these reference moving sources. The task of the subjects during the test was to always face the sound source. They sat on a swivel chair such that they could rotate freely without any interference from cables etc. Noise signals were randomly presented over loudspeaker or over the simulation. The procedure was the following: the sound source was located randomly somewhere in the frontal hemisphere for 4-6 seconds so that the subject had enough time to face the source. Then, the source began to move with a constant speed of 20 °/s for a certain time. In the end, the source stood still again for 4 seconds. In total, there were 32 trials preceded by 4 training trials to familiarize subjects with the test apparatus. The following conditions were used for the trials:

- 2 directions of movement (clockwise/counterclockwise)
- 4 different trajectory lengths: 30°/60°/90°/120°
- 2 conditions (loudspeaker/simulation)
- each trial was repeated once

The instructions were given in German, the original instructions can be found in Appendix B.

Prior to this actual test, the subjects were asked to do a short test to determine the sensor error. They had to face alternately the front loudspeaker (0° azimuth) and a given loudspeaker, in total 8 times with 8 different loudspeakers. The intention of this "calibration test" was to assess the "inherent" localization error of the subjects, that is, to determine a lower bound on the localization accuracy of the subjects. Unfortunately, we cannot separate between an error made by the subject and the sensor error. The sensor error is probably preponderant. To determine only the sensor error, one could attach a laser pointer to the sensor and ask the subjects to turn their head such that the laser pointer points exactly in the centre of a loudspeaker. With knowledge of the sensor error, the error caused by the test subjects could be estimated.

The whole test took $\sim$ 25 minutes. A retest was done after $\sim$ 1 month.

4.3.4 Results

From the recorded trajectories, we used two performance measures:

1. The RMS error during the moving phase of the sound source, defined by

$$RMS = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_{head}[k] - y_{source}[k])^2}$$

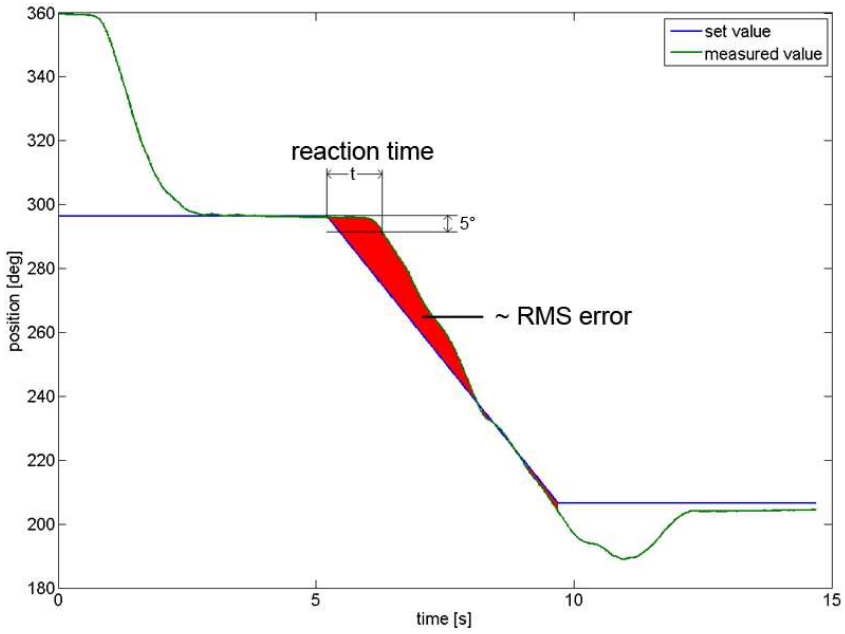


Figure 4.13: Source trajectory together with a typical measured trajectory and illustration of performance measures.

where y_{head} denotes the yaw angle of the head and y_{source} the yaw angle of the source.

2. The reaction time, defined as the time of the begin of movement of the source until the listener turns his head at least 5° in the right direction. This time is proportional to the MAMA, measured for that particular source velocity of 20 °/s at 0° azimuth.

Figure 4.13 shows a typical trajectory and illustrates the two evaluated criteria.

The mean localization error for the test and the retest are shown Figure 4.14. The corresponding numerical values are listed in Table 4.11. Significance was tested with a one-way ANOVA. There are significant differences between the mean RMS error of the simulation in the test compared to the retest, therefore, the test-retest reliability is poor. There are also significant differences between loudspeaker and simulated sources, but only in the test and not in the retest.

The mean reaction times for the test and the retest are shown Figure 4.15. The corresponding numerical values are listed in Table 4.11. Significance was tested with a one-way ANOVA. There are no significant differences between loudspeaker and simulated sources.

The results from the "calibration test", where the subjects had to face a certain loudspeaker, are shown in Figure 4.16. The errors were plotted in dependency of the traversed arc, e.g. if the subject had to face the loudspeaker

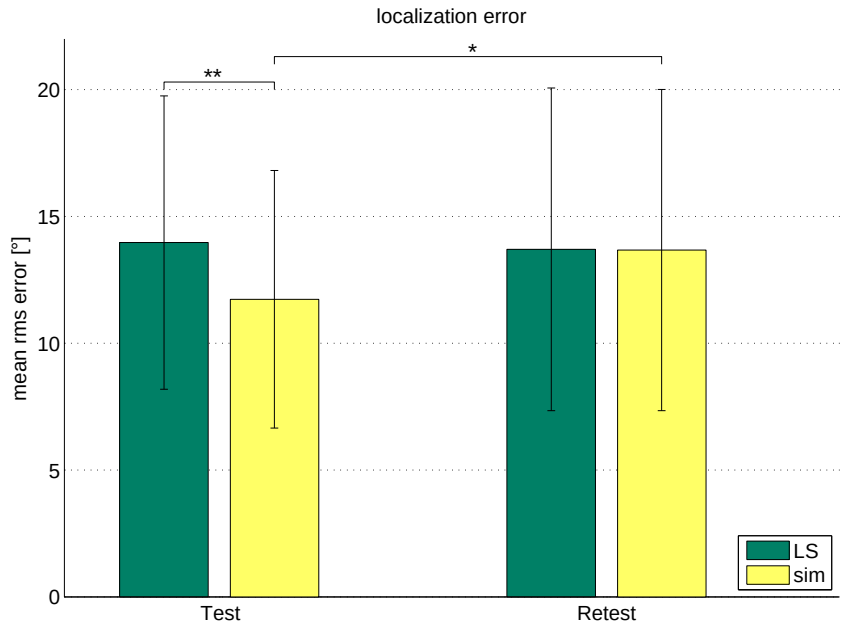


Figure 4.14: Mean RMS localization error of a moving source during the moving phase

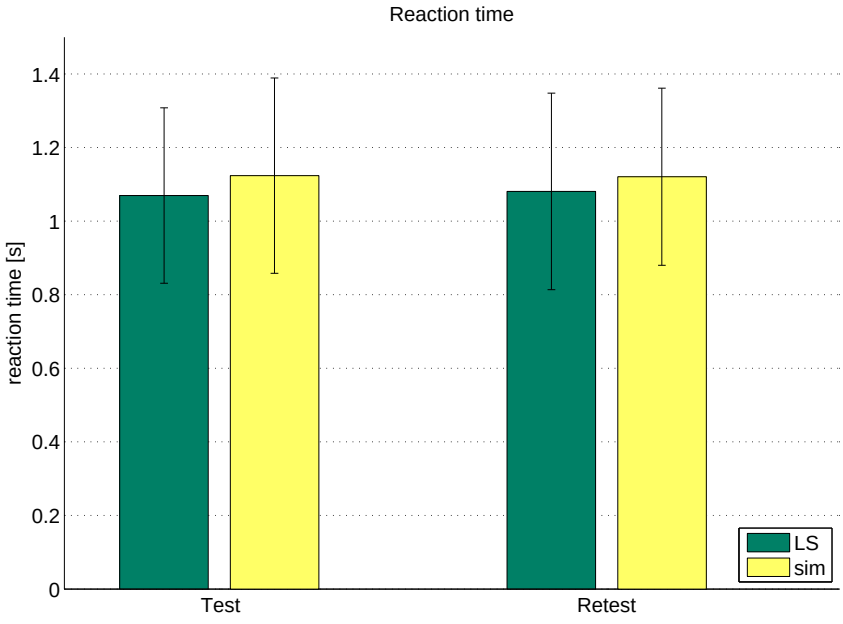


Figure 4.15: Mean reaction time until movement is detected

		Test	Retest	avg
RMS error	LS	14°	13.7°	13.85°
	sim	11.7	13.7	12.7
reaction time	LS	1.07 s	1.08 s	1.075 s
	sim	1.12 s	1.12 s	1.12 s

Table 4.11: Numerical results of the test with moving sources (mean values)

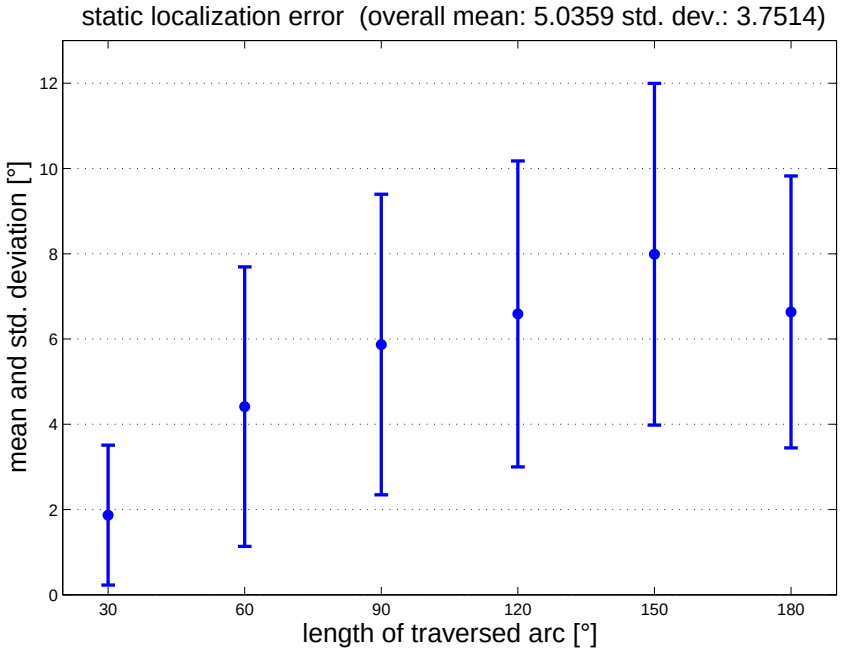


Figure 4.16: Mean static localization error in dependency of the traversed arc.

located at the left of the front speaker at 300° azimuth, the length of the traversed arc is 60°.

The results from the test with moving sources do not show any dependency on the length of the traversed arc nor on the direction of rotation.

4.3.5 Discussion

Sensor Error

The calibration test shows that the static localization error increases with increasing displacements from the front position. The error is maximal for a traversed arc of 150°, and the standard deviation is very large. There is no reason why the error made by the test subjects should increase with an increasing displacement from the front position, therefore the static error is most likely caused by the sensor inaccuracy.

RMS Localization Error

The results from the test with the moving sources were mixed. A big problem is the uncertainty about the sensor accuracy (cf. section 4.1). It is unknown, how large the sensor error in this test is and what the implications are. In the case of loudspeaker sources, the measurement of the head position is affected by this error. For the simulation, the head-tracker is used to measure the head position *and* to compensate head movements, with the result that the sound source is always positioned relative to the measured head position. In other words, the sensor error cancels out itself. To quantify the sensor error, one could attach a laser pointer to the sensor and let the subjects look at a specific location. However, this method does still not allow to quantify the sensor error in a dynamic condition.

Another unknown is the fidelity of the amplitude panning, which is used to present "moving" loudspeaker sources. A comparison with movable loudspeaker would provide clarity.

The mean localization errors that were measured were in the range of 12°-14°. To the best of our knowledge, nobody ever measured the localization *during* the moving phase, so the values cannot be compared to the findings of other researches. The results of the simulation and of the loudspeaker sources are roughly in the same range, but without knowledge of a) the impact of the sensor inaccuracy and b) the reason for poor the test-retest reliability with simulated sources, no conclusions can be drawn.

Reaction Time

The reaction time measurements were more consistent, in particular the test-retest reliability was good. The values from simulation and loudspeaker sources showed no significant differences. The mean values are in the range of 1.07 s to 1.12 s. With the chosen velocity of the sound source, 20 °/s, this corresponds to a minimum audible movement angle of 21.4° to 22.4°. The chosen definition of the reaction time, a rotation of the head of at least 5° in the correct direction, causes a systematic bias towards higher reaction times, since the rotation of the head by 5° takes also some time. If we assume that the subject turns his head at a rate of 20 °/s, then the bias is 0.25 s. This bias has to be subtracted from the measured reaction time.

In the literature, MAMAs of $\sim$ 8°-9° are reported for 0° azimuth. In contrast to our test setup, these values are determined by forced-choice procedures which are likely to be more sensitive. Together with the bias in our test, our results can be considered as plausible.

General Discussion

In summary, there are a lot of questions and only a few answers. For future listening tests with this system, a more accurate sensor is strongly recommended (cf. section 4.1). Then, it is important to know which property of the human auditory system is of interest and to design a test which is tailored to measure this property. A general test to study the localization of moving sound does not

exist. However, there are a lot of tests described in literature, all of them could be conducted with this system and the results could be compared with those of the literature. All these tests could also be conducted with hearing impaired subjects with or without any hearing aid algorithms.

Chapter 5

Conclusion

5.1 Conclusions

According to the title of this thesis, the main focus of this work would have been the efficient modeling of head movements and dynamic scenes in a virtual acoustics system. In reality, an existing virtual acoustics system was extended to present dynamic scenes and, more important, to compensate for head movements. Consequently, the focus of this work was set on the implementation and evaluation of the extended virtual acoustics system, whereas the efficient modeling was only a small part of it.

The existing virtual acoustics system provided a good basis to go one step further towards reality, since it is able to reproduce static scenes nearly indistinguishable from real scenes. This is the result from precise measurement of HRTFs, a room simulation software which is able to simulate rooms perceptually convincing by rendering both specular and diffuse surface reflections, and an open ITE speaker prototype which is able to reproduce the sounds naturally. However, the system is not tuned for the real-time rendering. It was intentionally designed as a flexible and portable system which runs on a standard PC with standard software. In contrast, the compensation of head movements requires a real-time rendering, which could be either realized by porting the system to specialized hardware, or by omitting some functionality which is not essential. The price of a system running on specialized hardware would have been the loss of flexibility and portability. The price of the chosen solution is a huge memory consumption caused by the offline rendered impulse responses, a restriction to horizontal plane movements and, most restricting, the support is limited to only one sound source. Apart from that, the new dynamic system extends the scope of application of the system to dynamic and even more realistic scenes where head movements are allowed to resolve front-back confusions.

The evaluation with subjective listening tests revealed that the compensation of head movements works very well. The untrained test subjects were not able to distinguish between simulated and real scenes reliably, although there are some weak indications that one could learn to detect the simulation. Apart from that, the stability of the impression was very good, while the externalization was not always fully convincing but still quite good. Listeners

who participated in earlier listening tests reported a subjective improvement of the externalization compared to the previous system. In particular, the head movement compensation led frequently to a persistent external impression of sound sources which were perceived in the head prior to a head movement. In summary, the extended system is more than a proof of concept. With a few extensions it could be used to assess the effect of different hearing aid algorithms on localization of sound source in realistic and complex scenes. The required extensions are described in the next section.

5.2 Future Work

The main drawback of the actual implementation is the restriction to only one sound source. A listening test in a complex scene (i.e., with background noise) is not possible with this system. The computational complexity increases linearly with the number of sources. The system currently uses only one CPU. If the calculations could be distributed across multiple CPUs, then it would be no problem to render also complex scenes with a multi-core processor. This should be possible with only a few changes in the code, since every CPU core could render one source – there are no data dependencies.

A very easy improvement of the system is the replacement of the sensor used in this work. The Xsens MTi sensor is fully compatible with the used MTx sensor, but the MTi does not show any drift, thus, off-position sound sources should be a thing of the past.

The last thing which is missing to test different hearing aid algorithms under realistic conditions, is the integration of these algorithms. In the existing system, the algorithms were simply applied to the static scenes prior to the playback. For the dynamic system, this approach is no longer possible since most of the algorithms are adaptive. A simple solution would be to use a real-time hearing aid simulation system, such systems exist, but they are rather expensive.

Finally, the listening tests have to be conducted. For static scenes, one could simply repeat the previous tests and allow head movements. For dynamic scenes, there is no all-in-one listening test. The test design depends heavily on the subject of interest. Moreover, there are some systematic mislocalization effects in dynamic scenes which have to be considered. Depending on that, there are several possible test setups, but further research in this direction is required.

There are also some ideas for an improvement of the system which are not essential, they fall in the category “nice to have”:

- Another interpolation scheme for the interpolation of the HRTFs could be evaluated. The scheme proposed by Ajdler [11] might provide a slight improvement over the used interpolation technique, resulting in more precise sound sources and/or to an improved externalization.

- The HRTFs could be measured for more positions and for elevations outside the horizontal plane. This might also lead to more precise sound

sources and/or to an improved externalization.

A last category includes improvements of technical nature. They could be implemented if the system should run on a slower computer or if computer resources are required for other tasks. The offline computation of a full set of impulse responses for one subject takes approximately one day and the resulting file has a size of 700 MB.

- The spatial resolution of the pre-rendered impulse responses could be reduced. The fine spacing of 1° is probably not really required. A spacing of 5° is too coarse, but a spacing of 2° could be sufficient and would require only half of the memory that the actual system is using.
- The reverberant tail of the impulse response could be replaced by a generic one. If the scenes are still perceptually convincing, one could save a lot of memory.
- The impulse response could be divided in more than two parts, which could be updated at a lower rate. This might result in a reduced computational complexity.
- The impulse response could be generated online. This would require a very powerful processor or the use of dedicated hardware. Such a system would have the advantage that arbitrary receiver movements could be compensated for.

Appendix A

Task Description

Master's thesis project proposal

Title:	Efficient modelling of head movements and dynamic scenes in virtual acoustics
Start date:	October/November 2009
Duration:	6 months
Location:	University Hospital Zurich (USZ), ORL clinic (ORL), Lab. for Experimental Audiology (LEA)
Supervisor:	Prof. Dr. Norbert Dillier

Introduction and motivation

LEA is currently involved in a project with Phonak titled "Hearing instrument algorithms for improved spatial perception". In the context of this project, we have developed the necessary hard- and software to virtually reproduce acoustics scenes over custom-build hearing instruments. At this stage, our virtual acoustics system is able to reproduce static acoustic scenes near-perfect and almost indistinguishable from reality.

The system is, however, currently unable to handle head movements of test subjects. If a subject moves his head, the virtual acoustic scene that is presented to the subject moves with the subject, rather then staying fixed in world coordinates. Furthermore, the current system is also unable to present dynamic scenes in which sound sources move around in space.

Previous research (for a recent study, see for example [1]) has shown that such head and source movements facilitate correct and accurate sound source localization judgments, reduce front-back confusions and allow better source elevation recognition. Similarly, we have found that the rate of front-back confusions is somewhat higher when virtual acoustics are involved compared to normal listening conditions. It is therefore desirable to extend our virtual acoustic system to include subject head movement and dynamic scenes.

Technical description

To simulate a static acoustic scene, our virtual acoustics system generates a binaural room impulse response (BRIR) for a given source and receiver position and orientation [2]. This involves computing the sound energy contributions at the receiver from specular and diffuse reflections of the sound source in the room's surfaces. All of these contributions, as well as the sound from the source that reaches the receiver directly, are convolved with a subject's individual head-related transfer functions (HRTFs) to model the effect of his head, torso and pinnae on the sound field. The use of individual HRTFs is essential for perceptually convincing virtual acoustics.

To simulate head movements and dynamic scenes, the BRIR must be constantly updated to reflect the new position of the sound source and orientation of the receiver. In our system, sound source position updates would come from a pre-described sound source trajectory, and receiver orientation updates would come in real-time from a motion tracker device. There are, however, a few limitations that need to be considered:

1. It is difficult if not impossible to generate the full BRIR in real-time with low latency.

   This problem is usually addressed by separating the BRIR into several parts, where each part is updated at a rate that is high enough so that the overall BRIR remains perceptually convincing, but that is low enough so that the entire BRIR is available in real-time and with low latency.

2. HRTFs are typically measured for a limited number of sound source positions on a sphere around the subject.

   The spatial sampling of HRTFs is usually dense enough such that convolving direct sound and surface reflections from arbitrary directions with their nearest-neighbour HRTF measurements will maintain perceptual accuracy. However, the spatial sampling may be too coarse to render small head movements and smooth sound source movements perceptually convincing. Typical solutions are to describe the measured HRTFs in a model, such that HRTFs for intermediate directions can be synthesized from the model, or to interpolate HRTFs for intermediate directions from adjacent measured HRTFs.

Project tasks and goal

The goal of the master's thesis project is to add the ability to deal with subject head movements and dynamic scenes to our virtual acoustics system. This would include but is not necessarily limited to the following tasks:

- Literature review to identify existing methods of dynamic virtual acoustics, and find fitting approaches to the limitations mentioned earlier.
- Setting up the motion tracking devices.

- Coupling real time data from the motion tracking devices to the virtual acoustics system.
- Dynamically updating acoustic parameters in the system to reflect head movements and moving sound sources.
- Modelling of HRTFs and/or perceptually accurate interpolation between adjacent directions.
- Perceptual evaluation of the dynamic virtual acoustics system.

References

[1] Wightman and Kistler, *Resolution of front-back ambiguity in spatial hearing by listener and source movement,* Journal of the Acoustical Society of America, 105(5), May 1999

[2] Steven M. Schimmel, Martin F. Muller, Norbert Dillier, *A fast and accurate "shoebox" room acoustics simulator,* Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp.241-244, 2009

Appendix B

Listening Tests

B.1 Original Instructions for Static Listening Test

Bei diesem Hörtest werden die englischen Sätze

> *She had your dark suit in greasy wash water all year.*
> *Don't ask me to carry an oily rag like that.*

von **verschiedenen** männlichen Sprechern gesprochen, die **Lautstärke variiert** dabei leicht. Das Signal wird entweder über einen Lautsprecher oder über die ITE-Devices abgespielt und wird so lange wiederholt, bis Sie die Wiedergabe unterbrechen. Vor jedem Versuch ist zudem noch eine Kalibrierung des Bewegungssensors nötig. Schauen sie dazu **exakt geradeaus** auf den Lautsprecher 9 und drücken Sie OK, wenn Sie bereit sind. Wenn das Signal abgespielt wird, dürfen Sie Ihren Kopf nach links und nach rechts drehen, aber vermeiden Sie es, den Kopf zu neigen. Bewegen Sie Ihren Kopf auch dann, wenn Sie die Schallquelle ohne Probleme orten können. Die Kopfbewegung ist also erwünscht, aber kein Muss.

Der Versuchsleiter wird Ihnen nach jedem Durchgang folgende Fragen stellen:

- Wo hörten Sie die Schallquelle? (Skala von 1-5, 1 = Im Kopf, 5 = aus dem Lautsprecher, siehe separates Blatt)
  Did you hear the sound source in your head or from the loudspeaker? (Scale 1-5, 1 = in the head, 5 = from loudspeaker)

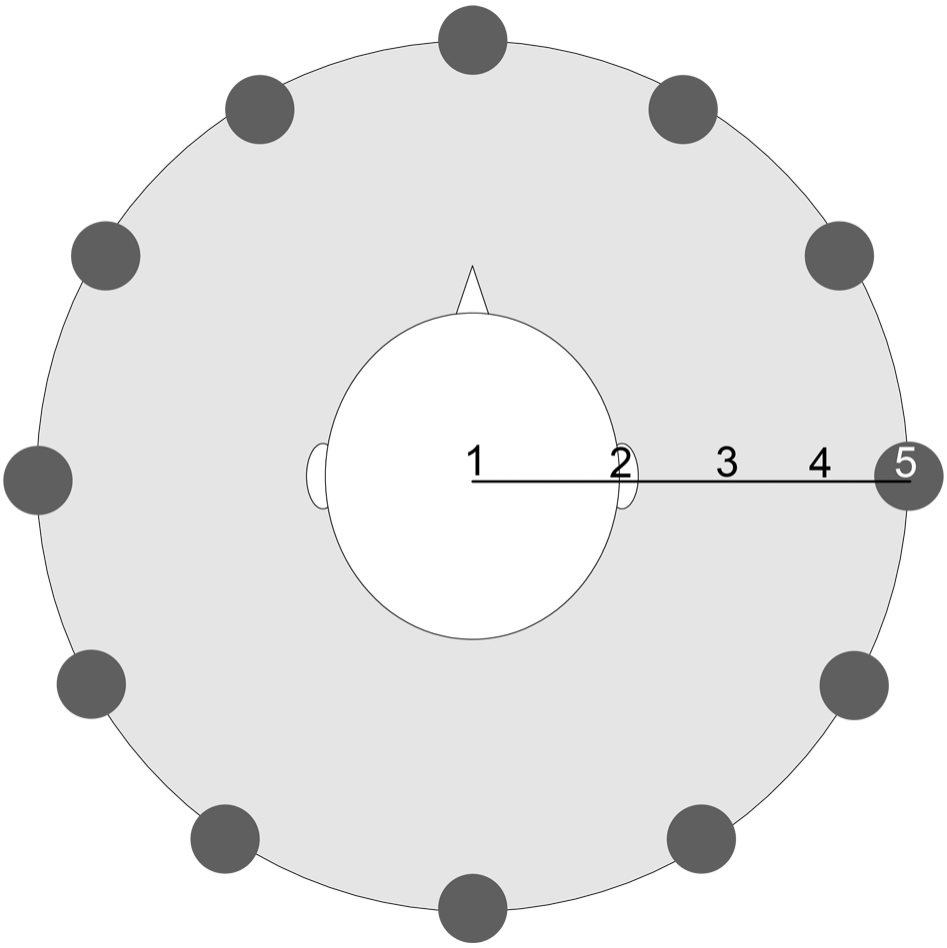
- Wenn Sie den Kopf bewegen, bleibt die Schallquelle an Ort und Stelle? (Skala von 1-5, 1 = unstabile Quelle, 5 = stabile Quelle, siehe separates Blatt)
  Does the sound source remains stable if you turn your head? (Scale 1-5, 1 = not stable, 5 = stable)

- Von wo kam der Schall: Lautsprecher oder ITE-Device?
  Where does the sound came from: Loudspeaker or ITE-Device?

- Was gab den Ausschlag für Ihr Urteil? Weitere Bemerkungen (offene Frage)
  Why? Remarks? (open question)

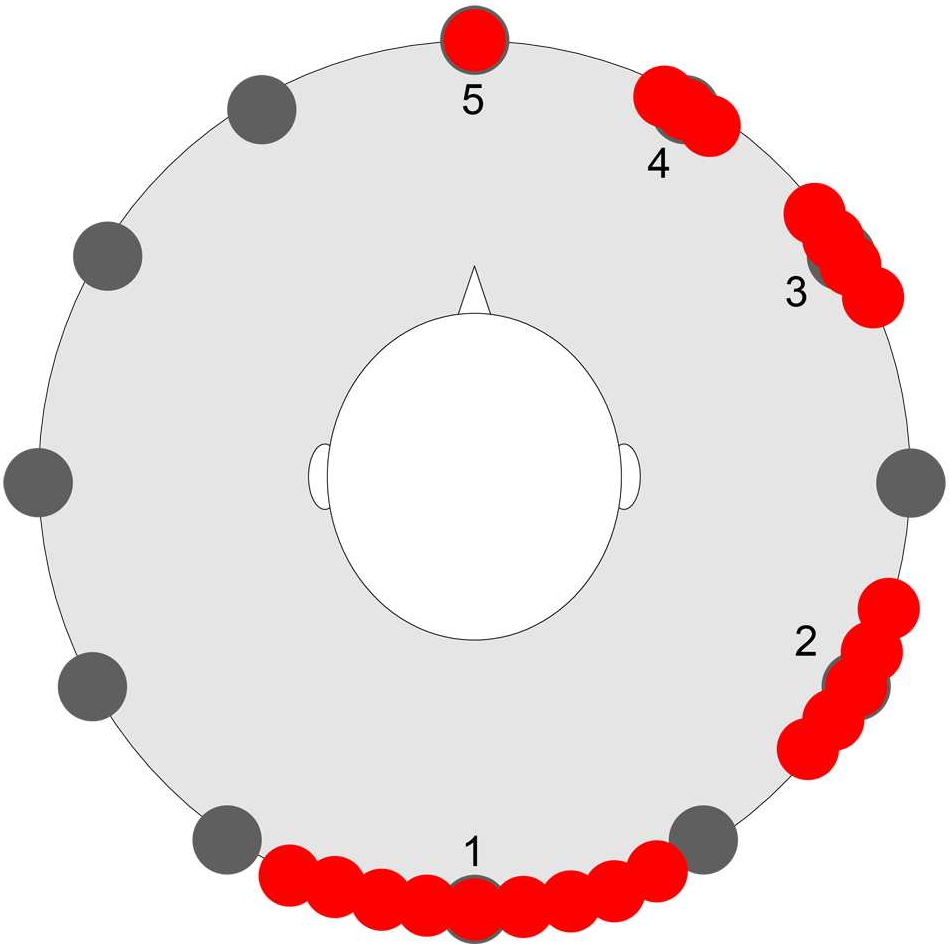
In einem zweiten Durchgang wird der gesamte Test wiederholt, aber dieses Mal mit verschiedenen Rauschsignalen.
Zu Beginn wird die Lautstärke der ITE-Devices mit der Lautstärke der Lautsprecher abgeglichen. Dann folgt ein Trainingslauf, um Sie mit dem Testablauf vertraut zu machen. Danach folgt der eigentliche Test.

Frage 1 (Wo hörten Sie die Schallquelle?)



Skala von 1-5:
1 = im Kopf
5 = aus dem Lautsprecher
2-4 gemäss Bild

Frage 2 (Wenn Sie den Kopf bewegen, bleibt die Schallquelle an Ort und Stelle?)



Skala von 1-5:
1 = völlig unstabile Quelle
5 = stabile Quelle
2-4 Zwischenstufen gemäss Bild

B.2 Proposed Instructions for Static Listening Test

During the static listening test, it turned out that the instructions are hard to understand because they contain too much information. There were no control questions to ensure that all test subjects fully understand their task. A draft for improved instructions (without changing the test procedure) is given here. The idea is to guide the subject step-by-step trough the instructions, such that he has more time to absorb all the information. This requires an interactive GUI. Buttons in the GUI are shown as a box, actions that should be executed after the user has pressed a button are marked with SMALL CAPS.

Screen 1

Herzlich Willkommen zu diesem Hörtest

Sie werden nun einen männlichen Sprecher hören, der die englischen Sätze

> *She had your dark suit in greasy wash water all year.*
> *Don't ask me to carry an oily rag like that.*

spricht. Die beiden Sätze werden so oft wiederholt, bis Sie die Wiedergabe unterbrechen. Wenn das Signal abgespielt wird, dürfesn Sie Ihren Kopf nach links und nach rechts drehen. Bewegen Sie Ihren Kopf auch dann, wenn Sie die Schallquelle ohne Probleme orten können. Die Kopfbewegung ist erwünscht, aber keine Pflicht. Drücken Sie auf `Start`, wenn Sie bereit sind

`Start`

PLAY SOUND FROM ANY LOUDSPEAKER (0°/90°/180°/270° AZIMUTH)

Screen 2

Der Sensor auf Ihrem Kopf zeichnet Ihre Kopfbewegungen auf. Er muss vor jedem Durchgang kalibriert werden. Dazu müssen Sie genau geradeaus auf Lautsprecher 9 schauen und `OK` drücken, wenn Sie bereit sind.
Versuchen Sie zudem, Ihren Kopf nicht zu neigen, wenn Sie Ihn drehen. Der Kopf sollte immer aufrecht sein.
Schauen Sie geradeaus auf Lautsprecher 9 und drücken Sie `OK`, wenn Sie bereit sind.

`OK`

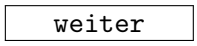
PLAY SOUND FROM EITHER LOUDSPEAKER OR SIMULATION (0°/90°/180°/270° AZIMUTH). SAME SPEAKER AS IN SCREEN 1

Screen 3

Der Versuchsleiter wird Ihnen nach jedem Durchgang folgende Fragen stellen:

- Wo hörten Sie die Schallquelle? (Skala von 1-5, 1 = Im Kopf, 5 = aus dem Lautsprecher, siehe separates Blatt)
- Wenn Sie den Kopf bewegen, bleibt die Schallquelle an Ort und Stelle? (Skala von 1-5, 1 = unstabile Quelle, 5 = stabile Quelle, siehe separates Blatt)
- Von wo kam der Schall: Lautsprecher oder ITE-Device?
- Was gab den Ausschlag für Ihr Urteil? Weitere Bemerkungen (offene Frage)

Schauen Sie geradeaus auf Lautsprecher 9 und drücken Sie `weiter`, wenn Sie bereit sind.



Play sound from either loudspeaker or simulation (0°/90°/180°/270° azimuth). Same speaker as in Screen 1. Test supervisor should ask the questions

Screen 4

Wenn Ihnen der Ablauf klar ist, drücken Sie auf `weiter`. Wenn Sie noch Fragen haben, hilft Ihnen der Versuchsleiter gerne weiter. Wenn Sie noch mehr Testläufe möchten, wählen Sie `mehr Testläufe`.

Continue with the test or repeat training (max. 2 additional training trials)

Screen 5

Control questions. To be defined.

Screen 6

Das Training ist nun beendet.
Es folgen nun 16 reguläre Durchgänge. Die beiden Sätze werden dabei von **verschiedenen** männlichen Sprechern gesprochen, die Lautstärke **variiert** dabei leicht.
Schauen Sie geradeaus auf Lautsprecher 9 und drücken Sie `weiter`, wenn Sie bereit sind.

`weiter`

16 PRESENTATIONS WITH SPEECH SIGNALS

Screen 7

Es folgen nun weitere 16 Durchgänge. Sie werden **verschieden gefärbte** Rauschsignale hören, auch bei diesen **variiert** die Lautstärke.
Schauen Sie geradeaus auf Lautsprecher 9 und drücken Sie `weiter`, wenn Sie bereit sind.

`weiter`

16 PRESENTATIONS WITH NOISE SIGNALS

Screen 8

Der Test ist nun fertig. Herzlichen Dank für Ihre Teilnahme!

`beenden`

FINISH TEST, RETURN TO MATLAB

B.3 Original Instructions for Dynamic Listening Test

Bei diesem Hörtest geht es um die Lokalisation von bewegten Schallquellen.

Als erstes wird der Kopfbewegungssensor auf Ihrem Kopf kalibriert. Dazu müssen Sie abwechslungsweise geradeaus und danach auf einen bestimmten Lautsprecher schauen (insgesamt 8 mal).

Im eigentlichen Test wird ein rauschartiges Signal aus den Lautsprechern oder aus den ITE-Devices abgespielt. Die Lautstärke und die Klangfarbe des Rauschens variiert dabei leicht. **Die Position der Schallquelle kann auch "zwischen" den Lautsprechern sein.** Ihre Aufgabe ist es, mit dem Kopf immer in die Richtung der Quelle zu schauen. Im Detail:

- Zuerst sollen Sie geradeaus auf den Lautsprecher 9 schauen
- Dann wird das Signal irgendwo in der vorderen Hemisphäre abgespielt. Drehen Sie Ihren Kopf in die Richtung der Quelle
- Nach einer kurzen Zeit beginnt sich die Quelle zu bewegen. Folgen Sie der Quelle, d.h. versuchen Sie, immer in die Richtung zu schauen, aus der der Schall kommt.
- Am Schluss bleibt die Quelle wieder stehen. Die Darbietung stoppt automatisch

Zu Beginn wird es vier Trainingsläufe geben, um Sie mit dem Testablauf vertraut zu machen. Dann folgen 32 Darbietungen aus den Lautsprechern oder den ITE-Devices.

Appendix C

Statistical Tests

C.1 Linear Regression Analysis

C.1.1 Simple Linear Regression

In statistics, linear regression is used to model the value of a scale variable y based on its linear relationship to one or more independent variables X (predictors). The linear regression model assumes that there is a linear, or "straight line," relationship between the dependent variable y and each independent variable x. The *simple* linear regression used in this work is the *least squares* estimator of a linear regression model with a single independent variable x. In other words, simple linear regression fits a straight line through a set of n points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible [24], [25].

Suppose there are n data points $\{x_i, y_i\}$, where $i = 1, 2, \ldots, n$. The goal is to find the equation of the straight line

$$y = a + bx$$

such that the line minimizes the sum of squared residuals of the linear regression model. The following minimization problem has to be solved:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

The solution to this problem is given by:

$$\begin{aligned}\hat{b} &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_i\sum_{j=1}^{n}\frac{y_j}{n}}{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2} \\ &= \frac{\overline{xy}-\bar{x}\bar{y}}{\overline{x^2}-\bar{x}^2} = \frac{\mathrm{Cov}[x,y]}{\mathrm{Var}[x]} = r_{xy}\frac{\sigma_y}{\sigma_x}, \\ \hat{a} &= \bar{y} - \hat{b}\bar{x},\end{aligned}$$

where r_{xy} is the sample correlation coefficient between x and y, σ_x, σ_y is the standard deviation of x and y, respectively. A horizontal bar over a variable denotes the sample average of that variable.

Substituting the above expressions for $\hat{a}$ and $\hat{b}$ into

$$y = \hat{a} + \hat{b}x,$$

yields

$$\frac{y-\bar{y}}{\sigma_y} = r_{xy}\frac{x-\bar{x}}{\sigma_x}$$

C.1.2 Coefficient of Determination

In statistics, the coefficient of determination, R^2, is the proportion of variability in a data set that is accounted for by the statistical model. The "variability" of a data set is measured through different sums of squares:

$TSS = \sum_i (y_i - \bar{y})^2$, the total sum of squares (proportional to the sample variance).

$ESS = \sum_i (\hat{y}_i - \bar{y})^2$, the explained sum of squares.

$RSS = \sum_i (y_i - \hat{y}_i)^2$, the residual sum of squares.

where $\hat{y}_i = \hat{a} + \hat{b}x_i$.

The ESS stands for the part of the variance in the data which is explained by the statistical model, whereas the RSS is a measure of the discrepancy between the data and an estimation model, i.e., the variance in the data which can not be explained by the statistical model.

Interpretation

In the case of simple linear regression, the coefficient of determination is defined as

$$R^2 = \frac{ESS}{TSS} = r_{xy}^2.$$

In this form, R^2 is given directly in terms of the explained variance: it compares the explained variance (variance of the model's predictions) with the total variance (of the data). R^2 is therefore a statistic that will give some information about how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly fits the data, while $R^2 = 0$ indicates that there is no linear relationship. An interior value such as $R^2 = 0.7$ may be interpreted as follows: "Approximately seventy percent of the variation in the measured data can be explained by the model."

C.1.3 Significance

The value $\hat{b}$ obtained by linear regression analysis may suggest that there is a positive or negative linear relationship between the independent variable x and the measured data y. However, it is an estimate and the linear dependency might also be the result of coincidence.

In order to calculate the significance usually one of the two possible assumptions is made: either that the errors in the regression are normally distributed (the so-called classic regression assumption), or that the number of observations n is sufficiently large so that the actual distribution of the estimators can be approximated using the Central Limit Theorem.

To determine if a linear relationship is *significant*, we use a statistical hypothesis test. First, we specify the null hypothesis:

$$H_0 : b_0 = 0,$$

which means that there is no linear relationship between x and y.

We test our null hypothesis with the so-called t-test:

$$t_{score} = \frac{\hat{b} - b_0}{SE_{\hat{b}}}$$

where $SE_{\hat{b}}$ is the standard error of the least square estimate $\hat{b}$. If the null hypothesis is true, t_{score} has a t-distribution with $n - 2$ degrees of freedom and is given by

$$t_{score} = \frac{(\hat{b} - b_0)\sqrt{n-2}}{\sqrt{RSS / \sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

It can be used to evaluate the probability that H_0 is true. Let p be the outcome of the t-test, then a value p close to 1 indicate that it is very likely that H_0 is true. For $p < 0.05$ or more strictly $p < 0.01$, the null hypothesis H_0 is rejected. The p-value corresponding to t_{score} can be found using a table of values from Student's t-distribution.

C.2 Analysis of Variance

The Analysis of Variance (ANOVA) is a common statistical method to test the hypothesis that the means of two or more groups are not significantly different. The method mainly compares the variance among and within the groups. The ANOVA assumes that the tested samples are independent, Gaussian distributed variables with the same variances [23], [24], [25].

If we consider an experiment where data from G different groups were acquired and described by their means $\mu_0, \mu_1, \ldots, \mu_{G-1}$, the goal of an ANOVA test is to see whether the means can be considered as belonging to he same distribution. We specify the null hypothesis

$$H_0 : \mu_0 = \mu_1 = \ldots = \mu_{G-1}.$$

If H_0 is rejected with a sufficiently high probability, it can be concluded that the means correspond to statistically different groups.

Let N_j be the number of observations in group j and X_{ij} the i-th sample of the group j, with $i = 1, 2, \ldots, N_j$ and $j = 1, 2, \ldots, G$, several basic statistical quantities can be estimated.

The mean of the j-th group is estimated as

$$\overline{X}_j = \frac{1}{N_j} \sum_{i=0}^{N_j-1} X_{ij}.$$

Similarly, the mean of all observations is estimated as

$$\overline{X} = \frac{1}{N} \sum_{j=0}^{G-1} \sum_{i=0}^{N-1} X_{ij}$$

where $N = N_0 + N_1 + \ldots + N_{N-1}$ is the total number of samples.

The ANOVA uses different sums of squares to estimate the global variability of the samples, their variability within one group and their variability between the groups.

$$SST = \sum_{j=0}^{G-1} \sum_{i=0}^{N_j-1} (X_{ij} - \overline{X})^2, \text{the total sum of squares.}$$

$$SSW = \sum_{j=0}^{G-1} \sum_{i=0}^{N-1} (X_{ij} - \overline{X}_j)^2, \text{the sum of squares within a group.}$$

$$SSA = \sum_{j=0}^{G-1} N_j (\overline{X}_j - \overline{X})^2, \text{the sum of squares among groups.}$$

The SST estimates the variation of each observation with the global mean $\overline{X}$, the SSW is an estimate of the variation of the samples within the G Groups

and the SSA estimates the variance of the means of all groups with respect to the overall mean. It can be shown that $SST = SSA + SSW$.

The sums of squares are not estimates of the variance σ^2, but σ^2 can be obtained by dividing SST, SSW and SSA respectively by $\frac{1}{NG-1}$, $\frac{1}{N-G}$ and $\frac{1}{G-1}$, which are called *degrees of freedom*.

Under the null hypothesis, we would still expect some random fluctuations in the means for the different groups due to the limited number of samples. Therefore, the variance estimated based on within-group variability should be about the same as the variance due to between-groups variability, which implies that their ratio is close to 1. If H_0 is wrong, this is no longer the case and the variability among the groups will be larger than the variability within the groups. A so-called F-test can be done on the estimate of the variance.

$$F = \frac{SSA(N-G)}{SSW(G-1)}$$

F follows a so-called F-distribution with $N-G$ and $G-1$ degrees of freedom and can be used to determine the probability that H_0 is true. If p is the outcome of such a test, a value p close to 1 indicates that it is very likely that the different groups of observations correspond to the same population. A value of $p < 0.05$ indicates a statistically significant difference, values of $p < 0.01$, or more strictly, $p < 0.001$ are considered as statistically highly significant. In bar plots, a line on the top of two bars is annotated with one ($p < 0.05$), two ($p < 0.01$) or three ($p < 0.001$) stars, corresponding to a statistically significant difference between the means of these two bars.

C.3 Wilcoxon-Test

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test for the case of two related samples or repeated measurements on a single sample. In contrast to the ANOVA, the Wilcoxon test does not rely on independent, Gaussian distributed variables with the same variance. Therefore, it is used if one of the assumptions the ANOVA relies on is violated (mostly the assumption that the data follow a Gaussian distribution) [24], [36].

The test is based on the comparison of the differences of related samples. In a first step, these differences are calculated. Then, the differences are ranked separately for positive and negative differences. Based on this ranking, a T-value is derived that can be compared to a table to determine the probability that the means of the two samples are equal.

Setup

Suppose we collect $2n$ observations, two observations of each of the n subjects. Let i denote the particular subject that is being referred to and the first observation measured on subject i be denoted by x_i and second observation be y_i. For each i in the observations, x_i and y_i should be paired together.

Assumptions

Let $d_i = y_i - x_i$ for $i = 1, 2, \ldots, n$.

1. The differences d_i are independent.
2. Each d_i comes from a continuous population (they must be identical) and is symmetric about a common median Z.
3. x_i and y_i are metrical or at least ordinal data, so the comparisons "greater than", "less than", and "equal to" are meaningful.

Test Procedure

The null hypothesis tested is: $H_0 : Z = 0$. The Wilcoxon signed rank statistic T is computed by ordering the absolute values $|d_i|$. To each ordered $|d_i|$, a rank number R_i in order of magnitude is assigned. In the case of two equal differences $|d_i|$, those differences are assigned the mean rank value. Let

$$\varphi_i = \begin{cases} 1 & \text{if } d_i > 0 \\ 0 & \text{if } d_i < 0. \end{cases}$$

If $d_i = 0$, the corresponding values x_i and y_i are discarded.
The Wilcoxon signed ranked statistic T_+ and T_-, respectively is defined as

$$T_+ = \sum_{i=1}^{n} \varphi_i R_i$$

and

$$T_- = \sum_{i=1}^{n} (1 - \varphi_i) R_i,$$

respectively. The final statistic T is given by

$$T = \min\{T_+; T_-\}$$

T is compared to a table of all possible distributions of ranks to calculate p, the statistical probability of attaining T from a population of scores that is symmetrically distributed around Z. A value of $p < 0.05$ indicates a statistically significant difference, values of $p < 0.01$, or more strictly, $p < 0.001$ are considered as statistically highly significant. In bar plots, a line on the top of two bars is annotated with one ($p < 0.05$), two ($p < 0.01$) or three ($p < 0.001$) stars, corresponding to a statistically significant difference between the means of these two bars.

Bibliography

[1] M. F. Müller, A. Kegel, S. Schimmel, N. Dillier, and M. Hofbauer "Localization of virtual sound sources in realistic and complex acoustical scenes", Technical Report, 2010.

[2] F. Wightman and D. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement", *Journal of the Acoustical Society of America*, 105(5):2841-2853, 1999.

[3] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP Journal on Advances in Signal Processing*, pp. 1-19, 2007.

[4] T. Takala, R. Hänninen, V. Välimäki, L. Savioja, J. Huopaniemi, and T. Huotilainen, "An Integrated System for Virtual Audio Reality", *100th AES convention,* 1996.

[5] G. M. Naylor, "ODEON — Another hybrid room acoustical model," *Applied Acoustics,* 38:131-143, 1993.

[6] K. Heutschi, "Skript Akustik I", Institut für Signal- und Informationsverarbeitung, Eidgenössische Technische Hochschule, CH-8902 Zürich, August 2007.

[7] S. Schimmel, M. F. Müller, and N. Dillier, "A fast and accurate shoebox room acoustics simulator," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 241-244, 2009.

[8] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[9] C. Moler, *Numerical Computing with Matlab*, Cambridge University Press, Cambridge, 2004.

[10] F. N. Fritsch and R. E. Carlson, "Monotone Piecewise Cubic Interpolation," *SIAM Journal on Numerical Analysis*, 17:238-246, 1980.

[11] T. Ajdler, C. Faller, L. Sbaiz, and M. Vetterli "Sound Field Analysis Along a Circle and its Applications to HRTFs Interpolation," *Journal of the Audio Engineering Society*, 56(3):156-175, 2008.

[12] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, 114(4):2236-2252, 2003.

[13] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, 27(4):339-368, 1940.

[14] S. Perrett and W. Noble, "The contribution of head motion cues to localization of low-pass noise," *Perception & Psychophysics*, 59(7):1018-1026, 1997.

[15] D. Perrott and A. Musicant, "Minimum auditory movement angle: Binaural localization of moving sound sources," *Journal of the Acoustical Society of America*, 62:1463-1466, 1977.

[16] A. Kulkarni, S. Isabelle, and H. Colburn, "On the minimum-phase approximation of head-related transfer functions," *IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, 1995.

[17] D. Kistler and F. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America*, 91:1637-1647, 1992.

[18] F. Christensen, H. Moller, P. Minnaar, J. Polgsties, and S. K. Olesen, "Interpolating between head-related transfer functions measured with low-directional resolution," *107th AES convention,* 1999.

[19] M. Matsumoto, S. Yamanaka, M. Tohyama, and H. Nomura, "Effect of time arrival correction on the accuracy of binaural room impulse response interpolation," *Journal of the Audio Engineering Society*, 52(1/2):56-61, 2004.

[20] V. Larcher, J.-M. Jot, J. Guyard, and O. Warusfel, "Study and comparison of efficient methods for 3D audio spatialization based on linear decomposition of HRTF data," *108th AES convention,* 2000.

[21] J. Chen, B. Van Veen, and K. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *Journal of the Acoustical Society of America*, 97(1):439-452, 1995.

[22] W. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy head microphone," Technical Report 280, MIT Media Lab, Perceptual Computing, 1994.

[23] M. F. Müller, "Speech Privacy in a medical application", Master Thesis, 2006.

[24] G. Clauss, F.-R. Finze, and L. Partzsch, *Statistik – Für Soziologen, Pädagogen, Psychologen und Mediziner,* Verlag Harri Deutsch GmbH, Frankfurt am Main, 2002.

[25] F. Brosius, *SPSS 11*, mitp Verlag, Bonn, 2002.

[26] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of the Audio Engineering Society,* 45(6):456-466, 1997.

[27] E. Zwicker and H. Fastl, *Psychoacoustics – Facts and Models,* Springer-Verlag GmbH, Heidelberg, 1990.

[28] D. Grantham, "Detection and discrimination of simulated motion of auditory targets in the horizontapl lane," *Journal of the Acoustical Society of America*, 79(6):1939-1949, 1986.

[29] K. Saberi and D. Perrott, "Minimum audible movement angles as a function of sound source trajectory," *Journal of the Acoustical Society of America*, 88(6):2639-2644, 1990.

[30] D. Perrott and A. Musicant, "Dynamic Minimum Audible Angle: Binaural Spatial Acuity with Moving Sound Sources," *Journal of Auditory Research*, 21:287-295, 1981.

[31] J. Middlebrooks and D. Green, "Sound Localization by human listeners," *Annual Review of Psychology*, 42:135-159, 1991.

[32] W. Gardner, "Efficient Convolution without Input-Output Delay," *Journal of the Audio Engineering Society,* 43(3):127-136, 1995.

[33] M. Frigo and S. Johnson, "The Design and Implementation of FFTW3," *Proceedings of the IEEE,* 93(2):216-231, 2005.

[34] R. Damgrave and D. Lutters, "The Drift of the Xsens Moven Motion Capturing Suit during Common Movements in a Working Environment," *Proceedings of the 19th CIRP Design Conference – Competitive Design,* 338-342, 2009.

[35] J. Garofolo, et al. "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Linguistic Data Consortium, Philadelphia,* 1993.

[36] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, 1(6):80-83, 1945.

[37] E. Macpherson and J. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *Journal of the Acoustical Society of America*, 111(5):2219-2236, 2002.

[38] J. Makous and J. Middlebrooks, "Two-dimensional sound localization by human listeners," *Journal of the Acoustical Society of America*, 87(5):2188-2200, 1990.

[39] [Online]. Available: http://www.howstuffworks.com/